AN ECONOMIC ANALYSIS OF ABILITY, STRATEGY AND FAIRNESS IN ODI CRICKET

A thesis submitted in partial fulfilment of the requirements for the Degree

of Doctor of Philosophy in Economics

in the University of Canterbury

by Scott R. Brooker

University of Canterbury

2010

Table of Contents

Acknowledgements

First and foremost, I wish to acknowledge my supervisor, Dr. Seamus Hogan, who always made time for me no matter how busy he was. I really appreciated being able to run downstairs and get Seamus' immediate thoughts on my latest work when on campus, along with rapid email responses at other times. Over the four years, Seamus' support has been outstanding and he found the perfect balance between keeping me on track and allowing me the freedom to control the direction in which I took the research. It is expected that a supervisor should enhance a student's understanding of the field, in this case economics, but Seamus also found the time to teach me about grammar, classic movies and the entire history of cricket prior to my birth in 1981. In summary, Seamus was the perfect person to supervise this project and for his efforts I am extremely grateful.

I would also like to acknowledge the support of those organisations who have contributed financially through fees scholarships, living allowances and/or conference funding: Sport and Recreation New Zealand (SPARC); The Tertiary Education Commission (TEC); the Economics and Finance Department and the College of Business and Economics Research Committee at the university. I reserve special mention for New Zealand Cricket (NZC) who have not only contributed financially to the project but also provided the majority of the data used in the thesis. In particular I would like to thank Dr. Peter Mayell for his efforts in coordinating my relationship with NZC in order to get the feedback of players, coaches and administrators, which helped to ensure that my models are in line with the way the sport is played.

Finally, I am grateful to the members of the Economics and Finance Department for their helpful suggestions, particularly during the various departmental presentations that I have given during the course of my research.

Abstract

The ground conditions prevailing on the day of a cricket match is an important confounding variable that results in the majority of cricket analyses requiring qualification. We present a Bayesian method for estimating the value of ground conditions in the absence of a direct measure. We use dynamic programming techniques to estimate models of both the first and second innings and we outline an application for each model. We extract a proxy variable for risk from our first-innings model and we use this variable to successfully estimate the trade-off between scoring rate and the probability of survival for individual batsmen. This enables us to decompose a batsman's performance into ability and strategic nous. Our second-innings model gives an estimate of a team's probability of winning at any point in the second innings of the match. We use this variable in conjunction with our ground-conditions variable to outline a new method for adjusting the target score in rain-affected matches. We introduce a simple metric for comparing the performance of various rain rules and we find that our proposed rule outperforms the incumbent Duckworth/Lewis method.

CHAPTER 1

Overview of the thesis

This thesis consists of four main chapters, distinct from one other by the nature of the theories and methods used, yet with each chapter possessing a close relationship with the other three chapters. It is intended that the thesis can be read as both a comprehensive piece of work and as individual chapters telling their own story. The purpose of this chapter is to discuss the nature of the individual chapters and explain their interdependencies: Later chapters depend on earlier chapters for their data needs, while earlier chapters depend on later chapters in order to properly justify their very existence.

Two ultimate goals of this thesis are to determine the abilities of various batsmen in One Day Internationals (ODIs) in a variety of game situations and to subsequently test their strategic optimality by assessing the likely outcomes of the decisions that they make concerning risk. Batsmen certainly have a variety of abilities and, given their ability, they have different strengths and weaknesses. This could be detailed down to the level of whether a batsman is proficient at playing a particular shot to a particular type of delivery; however, this thesis is primarily concerned with the rate at which a batsman is able to substitute between their scoring rate and their probability of survival. We show that this marginal rate of transformation is crucial in determining the optimal strategy for a given batsman in a given situation.

A separate goal is to assess the fairness of a selection of the various target-adjustment rules that have been used in ODI cricket over its history. We propose an alternative method and we show that our new method outperforms the incumbent Duckworth/Lewis method.

In any data analysis, there will be a set of variables that explain an outcome with some amount of random error. In some cases we do not necessarily have the values of all of these variables and if we are missing any important variables the analysis can be significantly distorted. This may be because the researcher is using a data set that was not collected particularly for their analysis or because the variable is very difficult to quantify, among other reasons. Cricket provides an excellent example of a variable that is difficult to quantify: the ease of batting due to the conditions.

On any given day, the weather and pitch conditions as well as the size and shape of the boundaries can have a significant influence on the size of scores that both teams are able to achieve. This is very important for our analysis as we need to separate out the variation of player and team scores caused by variation in skill from that caused by variation in conditions, in order to properly assess performance and strategy. Chapter 4 explains the reasons behind the difficulty in estimating conditions directly on any given day but a more pressing problem is that, even if such estimates are possible, they are not included in the data set obtained for the purposes of this thesis. In Chapter 4 we develop a method of indirectly estimating the value of this conditions variable. We use information from the distribution of first-innings scores and its relationship to the probability of winning to infer a Bayesian distribution of conditions for each match in our data set. This important variable is then included in our subsequent analyses in Chapters 5, 6 and 7.

In Chapter 5, we model the scoring rates and survival rates of batting teams in all possible game situations for the purpose of predicting the likely outcomes from any given state. We employ a dynamic programming approach by applying these modeled scoring and survival rates to the possible endpoints of the innings and solving by backwards induction until we have a complete set of predictions covering all possible game situations. A key variable in this modeling is the conditions estimate derived in Chapter 4, which enables us to predict outcomes assuming many different conditions.

Predicting the likely outcome from any given state of a match is interesting in its own right, but the largest contribution of Chapter 5 is that it provides us with a new variable that we can use to proxy risk-taking by the batsman, the cost of a wicket. Since the outcome from any game situation can be predicted by the model, it is also possible to assess the size of the negative impact of losing an immediate wicket on the likely outcome. We define this variable as the cost of a wicket and this information is extremely useful for the research in Chapter 6, which requires some knowledge of the risk intentions of a batsman.

Chapter 6 estimates production possibility frontiers (PPFs) for a selection of batsmen with the two "goods" being scoring rate (per ball) and probability of survival (per ball). On the surface, constructing PPFs are simple as it should be possible to look at our data and determine how often a player gets out when scoring at each rate and plotting these points on a graph. Upon closer inspection however, it is apparent that the probability of survival cannot be directly compared to the scoring rate in our data set. This is because, other than in rare cases where the batsman is run out having already completed at least one run, the batsman almost always scores zero runs from any ball that he is dismissed on. It therefore is the case that

running a regression of the binary variable *survival* on the continuous variable *runs* would lead to nonsense results as the number of runs scored from any ball where the batsman is dismissed is almost always zero.

The problem can be considered in an alternative way. A dismissal, to a batsman, represents an unsuccessful event. It is not possible to determine directly from the data set how quickly the batsman was attempting to score from the ball that just happened to lead to his downfall. He may have been playing some very aggressive shots or he may simply have been bowled a ball that was too good for him despite a defensive mindset, but all the data show is an outcome of zero runs, dismissed. To solve this problem it is necessary to consider the likely approach of the batsman as he faces each ball. He will be intent on taking a particular amount of risk each time the bowler runs in to bowl, but in the absence of data collected through a mind-reading device it is not possible to determine what the thoughts of the batsman were on any given delivery. The approach that we choose is to take the cost-of-a-wicket variable that was derived in Chapter 5 and use it as a proxy for the amount of risk taken by the batsmen. This has limited accuracy when considered ball-by-ball as batsmen do not necessarily engage in perfect strategy and may very well occasionally take high levels of risk when the cost of a wicket is very high and low levels of risk when the cost of a wicket is very low. However, when considering a batsman's behaviour on average over all the balls the he has faced, it is reasonable to assume that he has taken more risk when the cost of a wicket is lower. It is then possible to separately determine the relationships between the cost of a wicket and scoring rates as well as the cost of a wicket and the probability of survival. The predicted scoring rate and probability of survival for each value of cost of a wicket are then joined together to form the

points of the PPF for each player. As we are interested in a batsman's natural ability and strategic nous under a variety of ground conditions, the information from Chapter 4 is an important input into Chapter 6.

Finally, in Chapter 7 we develop a target-adjustment rule to be used in the event of match overs being lost due to rain. We construct a probability-preserving rule that we believe is conceptually superior to all previous such rules, with the largest improvement being the addition of a variable for the prevailing ground conditions, as calculated in Chapter 4. This enables the observed conditions on the day of a match to be specifically used as an input variable in the implementation of our rule. We show that our method compares favourably with the existing Duckworth/Lewis rule, which implicitly assumes that all variation in first-innings score is due to ground conditions.

CHAPTER 2

A review of the relevant literature

The statistical analysis of sports has rapidly grown in popularity in recent years, perhaps owing in part to the successful methods of Oakland Athletics general manager Billy Beane, who used previously unknown or unappreciated measures of performance to transform an under-resourced baseball team into one that was competitive with much higher-spending clubs. Beane's story was popularised by the book, *Moneyball* (Lewis, 2003), about his approach to the game.

Strategies in a variety of sports have been studied. Romer (2003) uses a dynamic programming approach to estimate the value of possessing the ball in different parts of the field in American football, to compare the expected payoffs from running or kicking the ball. Klaassen and Magnus (2008) calculate the optimal strategy for serving in tennis. Hirotsu and Wright (2003) apply dynamic programming to determine the best strategy for the configuration of a football (soccer) team, including the optimal strategy for making substitutions during a match. Dynamic programming can be used to estimate strategies designed not necessarily to score directly, but to put one in a better position to score, increasing the probability of winning.

Dynamic programming techniques are particularly suited to the analysis of cricket. The sequential nature of the game makes it, absent dynamic programming techniques, difficult to assess the current likelihood of each team winning, which in turn makes it difficult to

distinguish good strategies from poor ones. Cricket also has well-defined breaks and state variables, compared to sports that are controlled by the clock. This means that there are many non-arbitrary points of assessment. Cricket provides a rich array of data for a researcher to analyse; however, prior to the seminal paper on the topic of dynamic programming in cricket, by Clarke (1988), very little high-level research had been done. In particular, Clarke was surprised that Operations Research (OR) techniques had not been applied to cricket. He notes that "… lacking in the literature is the application of OR techniques to assist the cricketer with tactics. This seems strange given the role Britain and the Commonwealth have played in the origins and continued practice of both OR and cricket." It is likely that technological advancements, particular in the capacity and speed of computers, have played a large role in the recent increase in complex analysis of cricket.[1]

Clarke's seminal paper involves estimating a scoring rate and a probability of dismissal and he presents different models for the first and second innings. In the first innings the assumed goal of a batting team is to maximise its expected total score and in the second innings it is to maximise the probability of achieving the target score.[2] The model is set out under a dynamic programming framework, where the state space is divided into cells characterised in two dimensions by the number of balls and wickets remaining. Considering only the average ability of recognised batsmen, a relationship is assumed between scoring rate and dismissal rate based mostly on educated guesses, rather than detailed analysis. Clarke notes that estimates could be based on either expert opinion or statistical analysis and in addition, they should take into account the pitch conditions. In Chapter 4 we outline an analytical method for estimating a

[1] Clarke notes that he was programming in BASIC on an IBM XT.
[2] Maximising the expected total score is identical to our method, presented in Chapter 5, of maximising expected additional runs, as the past cannot be changed.

variable to represent ground conditions (including the pitch, outfield, weather and stadium size) and in subsequent chapters we show how the different conditions can affect a match.

Clarke uses his first-innings model to estimate the optimal scoring rate in any situation by using numerical simulation to investigate the outcomes under different scoring rates. His results suggest (pp. 333) "teams should try to score slightly faster than they expect their average rate for the rest of the innings to be, and if wickets are lost, slow up, rather than the current practice of scoring slower than average and speeding up if wickets are not lost. Thus the generally accepted view of scoring slowly at the beginning of the innings is not optimal under this model". Clarke also notes (pp. 334) that his model of the first innings can be used to compare various positions, such as "…is it better to be 1 for 50 or 3 for 80 after 25 overs?"

In Chapter 6, we build on this approach using substantially more detailed methods of calculating the scoring rate and the dismissal rate and we apply the notion of choosing an optional scoring rate to an individual batsman, rather than to a partnership as is the case in Clarke's paper.

Clarke's second-innings model is similar in formulation to his first-innings model, except that the variable to be maximised is the probability of winning, rather than the expected total. His conclusions from the first innings can also be generally applied to the second innings and he notes the existence of a second-innings advantage.

The paper also suggests that the probability of winning, calculated in the second-innings model, could be useful in the event of a rain interruption. In particular, Clarke uses an example to illustrate the shortcomings of the Average Run Rate (ARR) method. The paper stops short of

advocating probability maintenance as a criterion for determining revised targets, simply stating that it could be used to assess the performance of existing methods. In Chapter 7 we describe and assess a range of possible methods and outline a procedure where a probability-of-winning model could be directly employed in the target calculation.

In concluding, Clarke suggests that the dynamic programming model has many potential uses, such as quantifying the effects of including extra batsmen in the team, providing captains, coaches, commentators and even bookmakers with better measures of how teams are performing during the match, and developing measures of player performance that "better reflect the demands of one-day cricket" (pp. 336). Indeed Clarke went on to undertake many of these analyses himself or with co-authors, as well as applying dynamic programming methods to other interesting aspects of cricket such as the use of a night-watchman in tests. See Bailey and Clarke (2006), Allsopp and Clarke (2004), Clarke and Norman (1999), Norman and Clarke (2010), Clarke and Norman (2003) for some more specific applications of Clarke's dynamic programming work.

Preston and Thomas (2000) present a dynamic programming model similar to Clarke (1988); however, they impose a functional form for the probability of dismissal, which they call the dismissal hazard. Most relevant to this thesis is their inclusion of a dummy variable for each match in their dataset. This allows them to generate a maximum likelihood estimator of the dismissal hazard with a different intercept for each match, in order to control for differences in ground conditions and other innings-specific variables such as team ability. We believe that this approach is likely to attribute too much of the variance in the hazard rate to the innings-specific variables and not enough to the error term, emphasising the importance to nearly all cricket

analyses of having a reliable, impartial estimate of ground conditions. This is the motivation for our Chapter 4.

The results of Preston and Thomas suggest that, in contrast to the results of Clarke (1988), an approach of increasing aggression for a given number of wickets lost is the optimal strategy. They agree with Clarke that it is optimal to score at above the run rate required in earlier partnerships in the second innings.

The two optimal strategy models presented so far have assumed some kind of relationship between scoring rate and dismissal rate. Clarke employed "guesstimates" of the relationship, and Preston and Thomas estimated the relationship by first assuming a functional form. Barr and Kantor (2004) plot the performances of individual players in scoring rate / dismissal rate space and assume a functional form for "curves of equal suitability" (pp. 1269), akin to indifference curves. There are two important limitations of their method. First, the indifference curve is not based on any particular analysis of what the optimal strategy should be, and second the performance of each player is represented by a single point. They compare their analysis to risk-return models used in financial analysis, noting the limitation that "…while optimal combinations of assets with attractive risk-return characteristics can be combined to form efficient frontiers in financial risk-return space, the batting characteristics of people cannot be combined" (pp. 1268).

In Chapter 6 we outline a method of effectively creating frontiers in risk-return space: specifically, estimating the Production Possibility Frontier (PPF) of any batsman, where the scoring rate and survival rate (the complement of the dismissal rate) represent return and risk, respectively. These two components are estimated separately as functions of a proxy variable

for risk. This means that the assumption of a specific functional form for the relationship between scoring rate and survival rate is unnecessary, resulting in a much more flexible function.

A large fraction of the academic literature on cricket analysis has been devoted to the question of the fairest way to adjust the second-innings target score in a match shortened because of weather. Papers in this literature include Duckworth and Lewis (1998, 2004, 2005), Preston and Thomas (2002), Jayadevan (2002), Carter and Guthrie (2004, 2005) and Manage *et al.* (2010). The Manage *et at* paper uses Receiver Operating Characteristic (ROC) curves to assess the accuracy of the different rain rules. A common factor in their approach and our analysis in Chapter 7 is the method of creating artificial abandonments in fully completed games in order to assess the predictive power of each rule. The ROC method plots the sensitivity (true positive rate) versus the specificity (1 - false positive rate) in order to determine the trade-off between the two. Their paper gives two examples where the ROC curves are used: umpiring decisions and rain rules. While we find their method interesting, we do not believe that this is the most appropriate way to assess different rain rules, preferring instead to calculate a simple Correct Prediction Percentage (CPP) as outlined in Chapter 7.

The four main chapters of this thesis are substantial improvements on the existing models. A variable for ground conditions, calculated in Chapter 4, would be an excellent addition to most of the models in the existing literature. We use the ground conditions variable to create more accurate dynamic programmes to estimate first-innings future run-scoring and the probability of winning in the second innings in Chapters 5 and 7, respectively. We develop an advanced estimation strategy for determining the PPF of a batsman in scoring rate and

survival rate space, which crucially allows us to divide his performance into two components – ability and strategic nous. Finally, we propose a new rain rule that takes into account the ground conditions and assess the performance of our rule against a selection of previously used rules in Chapter 7.

CHAPTER 3

A description of cricket in an economic framework

3.1 Introduction

In some countries, it is almost impossible for a child to grow up without gaining at least a basic understanding of the game of cricket, even if they have never played the game. Despite this, while cricket is immensely popular in many countries, there are many parts of the world where cricket has not yet achieved a strong following and the nuances of the game can be a complete mystery to those who are not cricket fans. This short chapter has two goals. First, in Section 3.2 we provide a brief description of the game, containing the necessary rules and strategic knowledge required for proper understanding of the chapters to follow. This section should be unnecessary for readers who are familiar with the game of cricket. Second, in Section 3.3 we outline how this thesis will conceptualise the essential components of cricket as a strategic economics game. This provides an insight into the thinking behind the way that we model the game in subsequent chapters.

3.2 The game of cricket and our necessary assumptions

Cricket is a two-innings sequential game where one of the teams (Team 1) bats first, attempting to score as many "runs" as possible. It is then Team 2's turn to bat and their goal is

to beat the score achieved by Team 1. Note that a coin toss at the start of the game determines which team will bat first. There are many forms of the game of cricket; in this thesis we focus on One Day International (ODI) cricket, where each team faces a maximum of 300 legitimate balls during their innings. A team's innings has two constraints - the 300-ball limit and a maximum of ten wickets (outs). When either of these constraints is reached, or in the case of the second innings the target score is achieved, the innings is terminated. Each batsman bats only once during a match and the game requires that two batsman be batting at any one time, which becomes impossible once 10 of the 11 players are out.

Before we can consider creating a model of the game of ODI cricket, we need to set out the framework upon which our model will be built. It is important to identify the different stages of the game, define the variables involved in determining the outcomes and make the necessary simplifying assumptions that enable the model to be efficiently built.

3.2.1 The stages of a One Day International match

An international team usually has at least 12 players present at the venue of the match, from which 11 are chosen to play in the match. We treat the available players as fixed as the squad has generally been selected to play a series rather than an individual match. Following this assumption, there are 603 stages in a One Day International, as outlined in Table 3.1.

Table 3.1: Stages of the Game

Stage Number(s)	Description of Stage
1	The naming of the teams
2	The decision at the toss
3 - 302	Balls 1 to 300 of the first innings
303-602	Balls 1 to 300 of the second innings
603	The declaration of the result

3.2.2 The decisions required at each stage

A One Day International begins with the naming of the teams (Stage 1). This must be determined before the coin toss and therefore constitutes its own separate stage. A captain will select 11 players from his available squad based on the relative strengths and weaknesses of his squad and the opposition squad, player fitness, weather conditions and pitch conditions among other factors. Another consideration is the bat-first or field-first decision that he wants to make if he wins the toss, weighed against his expectation of the opposition captain's decision if the outcome is reversed. A captain is unlikely to want to select such a team that would be severely weakened if the result of the toss goes against him.

Once the coin toss has taken place, the game is in Stage 2. The captain winning the toss must decide if his team's interests are best served by electing to bat first or field first. The same factors from Stage 1 that could influence his decision apply again.

Stages 3 to 602 constitute the actual playing of the game - Stages 3 to 302, the first innings, and Stages 303 to 602, the second innings. Before each ball is bowled three decisions are made: The fielding captain decides where he wants to place his fielders, acting within the

restrictions imposed on him at the time;[3] the bowler decides the type of delivery that he is going to bowl to the batsman; and the batsman facing the next ball decides how aggressive he wants to be.

The first two decisions are based on the expected amount of risk-taking of the batsman and the desired amount of risk for the bowling side. For example, the fielding captain may suspect that the batsman is very aggressive and is going to try to hit the ball out of the park for six runs. In response to this, he may position as many fielders as possible on the boundary and instruct the bowler to bowl a ball of full length, as it is difficult for the batsman to get his bat underneath this type of delivery. This would be the fielding captain's best chance of preventing a six. However, the fielding captain might be willing to concede a six in exchange for an increased likelihood of getting the batsman out, in which case he may position his field and instruct his bowler differently.

The batsman's decision should depend on the game situation, the strengths and weaknesses of the bowler he is facing and the placement of the fielders (who must already be positioned before the bowler runs in to bowl). The batsman's approach is most likely a set of conditional behaviours, based on all the types of ball that he could get from the bowler he is facing. In the example above it was the case that the batsman would be prepared to try to hit every ball with maximum power. We often see this behaviour towards the end of an innings, where wickets tend to be relatively less valuable as the 300-ball limit becomes the constraint more likely to end the innings. In other game situations, the batsman may be intent on

[3] The laws of cricket restrict the fielding captain from placing his nine available fielders (it is clearly defined where two of the fielders, the wicket-keeper and the bowler, will be positioned) in certain combinations. These restrictions are not constant throughout an innings.

defending balls which are well-bowled and only attempting to hit the poor balls with power. This decision regarding the amount of risk to be taken will be shown to be critical to the work to be presented in subsequent chapters.

In Stages 3 to 602 there are a number of decisions which are made periodically by the batting and fielding captain, although not at every stage. In Stage 3 the batting captain must decide which two members of his team will be the first two batsmen and the fielding captain must decide which member of his team will bowl the first over (a set of six balls). The fielding captain must repeat this decision after every six balls and he is constrained by the restriction that each bowler may bowl a maximum of ten overs. The batting captain must make a decision at the fall of every wicket other than the ninth or tenth wicket, as he must decide which player from his team will be the next batsman. There is no decision to be made at the fall of the ninth wicket as there will only one batsman who has not yet batted and at the fall of the tenth wicket there is no more batting to be done. Generally speaking, the first five or six batsmen are of similar ability and are ordered according to the strategic preference of the captain, while the ability of the remaining batsmen is lesser for every drop in position in the order.

Finally, Stage 603 is a simple stage containing the result of the match, after the final ball has been bowled. This result, from the perspective of either team, can be a win, a loss, a tie or a no-result. Note that these latter two, which both generally result in any league table points from the game being shared, are substantially different outcomes. A tie occurs when the game reaches its natural conclusion with Team 2 having scored exactly one fewer run than its target score. A no-result simply indicates that there was simply not enough cricket played to determine a winner. This would normally be because of poor weather making it impossible to

complete the allotted number of balls in each innings. Poor weather causing a reduction in the time available to complete the match, however, does not necessarily result in a no-result, as mechanisms exist to adjust Team 2's target score to compensate for the reduced time. We discuss these in Chapter 7.

3.3 Cricket decision-making in an economic framework

From the first ball of the first innings to the last ball of the second innings, cricket is essentially a sequence of 600 two-move sequential games. In each of these games, the fielding team presents a frontier with a trade-off between risk and return to the batsman at the crease. They do this by selecting the positioning of the fielders and attempting to bowl the ball in a particular area of the pitch. The batsman selects a point on that frontier by choosing a level of aggression, before nature stochastically determines a realisation from the chosen point in terms of runs and/or a wicket.

There are two important components to the determination of the outcome from a given ball. First, batsmen, bowlers and fielders have varying natural abilities and it is these abilities in combination that determine the range of frontiers that a fielding side can present to a particular batsman. Second, the players must make strategic choices in the frontier that the fielding team presents to the batsman and the point along the frontier that is chosen by the batsman.

Throughout this thesis we focus our attention on the risk level chosen by batsmen, rather than the risk level chosen by bowlers and fielders. The overall amount of risk is a function of the risk level selected by each team; however, we show in Chapter 5 that it is the

batting team that has substantially greater control on the overall level of risk. As a result of this we predominantly choose to model the game of cricket as if bowling simply involves the execution of physical skills while batting also involves strategic choice.

CHAPTER 4

Inferring the ground conditions: A Bayesian approach

4.1 The influence of ground conditions

The ground conditions present on the day of a match play an important role in the sport of cricket, but they are certainly not easy to measure. In this chapter, we develop a method of inferring the contribution of the variability of ground conditions to the total variability of the first-innings scores. We use this information to construct probability distributions for ground conditions conditional on the observed score and outcome of each match. Our process enables us to extract from the data a measure of ground conditions where no direct measure is available in our data set.

The outcomes that take place on a sports field are closely related to the performance and ability of the players or athletes taking part in the sport; however, these are not the sole determinants. The sporting world contains many examples where factors unrelated to player and team ability have an impact on the type of game played and the result. The main factors of this type relate usually to weather conditions either prior to or during the time of competition, as well as the characteristics of the venue. The impact varies significantly from sport to sport. In sports mostly played indoors, such as basketball, the impact of weather conditions should be close to zero but "stadium" factors such as the quality of the lighting may have an impact on the level of scoring. In rugby union, wet and muddy conditions often lead to a more

conservative game, involving less lateral movement of the ball and lower scores. In sprinting, athletes are able to run faster with the wind, but the administrators of the sport do not recognise a world record time if there is deemed to have been substantial wind assistance. In the extreme, a sporting event may not even take place because of weather conditions; for example, sailing events may be postponed due to insufficient wind.

In cricket, there are five main factors that influence the first-innings score as well as the likelihood of each score being a winning one. These factors are

- the skill levels displayed by the players on both teams;
- luck;
- ground size;
- pitch conditions; and
- weather conditions.

The skill measure has two components. The overall strength of the teams as well as their relative strength in bowling, fielding and batting all have an influence on the likely score. The second component to the skill measure is the actual performance on the day of the two teams, given their overall strengths in each discipline of the game.

Luck plays a role in the outcome of a match; for example, poor umpiring decisions can have a marked influence, as can uncontrolled aerial shots that fall safely rather than going directly to the fielder.

On a small ground, it is relatively easier for the batsmen to hit the ball out of the playing field for boundaries and for this reason scores tend to be higher on small grounds than on large grounds. A mitigating factor here is that there are generally fewer twos and threes run as batsmen more often have to settle for single runs due to the ability of the fielders to reach the

ball faster on a smaller ground. A fielding side should, however, be at a minimum indifferent if they were given the option to change from a small ground to a larger ground, as the larger ground simply creates more options for possible field settings, as well as making it more difficult for the batsmen to hit boundaries.

Pitches are extremely variable in their nature. The moisture content, the type of soil used, the hardness, the amount of grass and any cracking present on the pitch all have an impact on how the ball behaves when it bounces on the pitch. Any movement or change of direction of the ball after hitting the pitch makes batting more difficult, as does inconsistent bounce, extreme pace off the pitch and extreme lack of pace off the pitch. Pitches are very individual; therefore, it is not appropriate to assume that all pitches at a particular ground will behave in the same way.

A fascinating aspect of the game of cricket is the tendency of the ball to “swing”, or change direction, in the air after it has been bowled. This swing, if present, makes batting significantly more difficult and is likely to lead to lower scores. On a cloudy or humid day the ball generally swings significantly more than on sunny dry days. For this reason the weather is our final factor influencing the outcome of the game.

It is useful to categorise these factors into two groups, based on the degree to which they are the same for both teams on any given day. The skill level is clearly team-specific and luck should be completely random; therefore, we combine these factors into a category entitled “performance”. The size of the ground obviously does not change during the game, and while pitch and weather conditions might change somewhat over the course of a match, we assume

that these factors vary to a far lesser degree within a match than between separate matches. We assign these three factors to a category entitled “conditions”.

4.1.1 The danger of ignoring the influence of conditions

Our analysis of the game of cricket is limited if we ignore the variability of ground conditions. We explain this by way of example. One of the models that we outline in Chapter 5 predicts the average additional runs scored from any possible situation in the first innings. If we do not include a variable for ground conditions in our model, we are effectively assuming that all ground conditions are the same. It seems intuitive that this would be a model for what would happen in average ground conditions, but on closer inspection this is not the case.

Consider a team that makes a very poor start to a match, perhaps two batsmen are out on the first two balls of a match. Given this start, it is more likely than not that this match is being played in worse than average ground conditions, from the point of view of the batting team. This likelihood, implicitly built into the model, means that the predicted average additional runs for this situation will incorporate the fact that we have a higher probability of being in poor batting conditions than good batting conditions. If we are, in fact, on an average pitch and the poor start was due to bad batting, good bowling or simply luck then our model is going to underestimate the expected number of future runs. The opposite holds for situations where the batting team makes a very good start. In this situation they are more likely than not to be playing in better than average ground conditions. If conditions on the day are in fact average, meaning that the good start is due to factors other than conditions, we will overestimate the expected additional runs that can be scored in these average conditions.

Effectively, ignoring ground conditions in our model means that we are estimating reduced-form coefficients. In order to construct a predictive model of what would be expected to happen under different ground conditions, it is important to include a ground conditions variable in our model to enable the estimation of structural-form coefficients.

There is an important difference between a situation where all ground conditions are the same and a situation where we know that ground conditions may vary but we are ignorant of their variations. In the former case we would not require a variable for ground conditions and our estimated model would obviously not be biased by this omission. In the latter case if we proceed with our modeling without including a variable for conditions then our model will make implicit assumptions about conditions based on the strength of the position that the batting team is in at any given time. Including a ground conditions variable in our model has two effects: improving the accuracy of the model for any particular ground conditions and identifying the sensitivity of the model in various situations to different ground conditions. It is important when constructing a strategy to know how much you should adapt your strategy to various conditions and the optimal adjustment is unlikely to be constant throughout the game.

Duckworth and Lewis (2005) are critical of the proposed target adjustment method of Carter and Guthrie (2004), stating that they do not take ground conditions into consideration. The model proposed in Duckworth and Lewis (1998), implicitly assumes that all variation in first-innings scores is due to variation in ground conditions. We note that the difficulty of estimating ground conditions is likely to have forced researchers to adopt either one of these extreme points of view, while the truth is likely to be somewhere in-between.

4.1.2 Why is a ground conditions variable missing?

Directly observing a ground conditions variable is extremely difficult. While the size of a particular ground is generally constant, weather conditions and the nature of the pitch are certainly not. Some grounds are more likely to have certain weather and pitch conditions than others, but there is significant variation due to the time of year and beyond this a large random component. Measuring the pitch conditions would be extremely difficult and measuring the effect of the weather conditions would be almost impossible, even given historical weather records, as there is such a variety of factors that can make a ball swing. There may be, to the naked eye, two identical days and the cricket ball may swing one day and not the next.

More importantly, even if a cricket-expert observer could estimate a value for ground conditions on the day of any given match, it is not easy for a modeller to objectively obtain ground conditions data. Furthermore, our data set is historical and therefore determining the nature of the pitch in particular, in games that were in some cases played several years ago is very problematic.

In light of both the natural difficulty of determining ground conditions and their omission from all known historical data sets, we decide that an indirect approach to estimating the ground conditions is required. In the remainder of this chapter we present a possible approach.

4.2 The simple theoretical model

We create two variables by separating the factors influencing the outcome of a game into two groups. Those factors that are specific to the ground conditions on the day are ground size, pitch conditions and weather conditions. We combine these factors into a variable "Conditions". The remaining factors, skill and luck, ought to be independent of the ground conditions on the day and we combine these factors into a variable "Performance".

For ease of interpretation, we consider performance to be a positive function of the batting team's skill and luck and a negative function of the bowling team's skill and luck; therefore, an above average value for performance will on average imply a better performance by the batting team than the bowling team, but not a particular level of either batting or bowling performance.

We consider the value of conditions to be the expected value of the number of first-innings runs that the average batting team would score against the average bowling and fielding team in the prevailing conditions on the day.

Let S be the first-innings score, ρ be the measure of "Performance" and χ be the measure of "Conditions". Given that a game contains two innings, define the performance of the team batting first (Team 1) as ρ_1 and the performance of the team batting second as ρ_2. We assume that both teams face the same conditions, so χ is constant throughout a match. The winning team is the team with the greater ρ. We define the relationship between score, performance and conditions as

$$S = \rho + \chi$$

Ideally we would have data for χ, but absent this information we want to construct an estimate for χ which is based on the observable information. If we knew the distributions of S, ρ and χ as well as the probability of winning given the value of these three variables, then we could apply Bayes' rule to construct a posterior distribution of χ given the first-innings score and the result. We would then have an estimate of conditions constructed solely from observable information. By assuming that S, ρ_1 and χ are normally distributed, and assuming that ρ_2 is drawn from the same distribution as ρ_1 but has a constant added to represent the second-innings advantage, the only information that we are missing are the means and variances of the three normal distributions. The mean and variance of S can be calculated easily and we have already defined χ as the average score that will be achieved when two average batting and bowling teams play each other; therefore, the mean of χ is equal to the mean of S and the mean of ρ_1 is equal to zero. All that we are missing is a decomposition of the variance of S into two parts, representing the variance of ρ_1 and χ. With this information, we would be able to estimate the posterior distribution for conditions.

We assume an independent, additive relationship between our two right-hand-side variables as we do not expect the deviations (in terms of number of runs) from the value of conditions of the total scores achieved to vary between different sets of conditions. In other words, we expect a score that is α runs in excess of χ to be equally competitive for a given value of α, independent of the value of χ.

We further assume that ρ and χ are normally-distributed variables having distributions $\rho \sim N(0, \sigma_\rho^2)$ and $\chi \sim N(\mu_\chi, \sigma_\chi^2)$, where $\mu_\chi = \mu_S$. In order to create a tractable model, it is convenient to assume that ρ and χ are normally distributed as this implies that *S* is also normally distributed. Our approach can nevertheless be justified by the application of the central limit theorem to an understanding of the game of cricket.

The performance measure is a combined measure of batting team performance and fielding team performance. The batting team performance is composed of the individual performances of up to 11 batsmen and the fielding team performance is composed of the individual performances of up to 11 bowlers and fielders. Each player may not play an equal part in determining the overall performance of the teams, but generally speaking the central limit theorem would imply that there are more ways of putting together the 22 performances in a way that gives an average overall performance than there are ways of putting them together to get an extremely good or extremely poor performance. Furthermore, with 300 individual balls in an innings, performance will also vary from ball to ball even within the overall performance of an individual player. The most extreme performances would require an extremely good performance from all required members of one team and an extremely poor performance from all required members of the other team. This would be much less likely than an average total performance, which could be caused by almost unlimited combinations of good batting and bad bowling from various players, or vice versa, completely cancelling each other out. This is true even if the individual player batting and bowling performance distributions were uniform.

We can make a similar argument for the normality of the conditions distribution. Conditions are a combination of a number of individual factors such as the nature of the pitch,

ground size and weather conditions. These main factors are likely to have smaller factors underpinning them, with each sub-factor requiring a draw from a distribution for each match. It is, however, not as obvious that our conditions distribution should have a normal distribution as it is for our performance distribution, due to at least some factors, such as rainfall and soil type, being relatively constant at a particular venue or at least correlated with a particular country. Later in the chapter we show that normality is a reasonable assumption for *S*, which increases our confidence in the normality of χ .

Since the sum of two normally distributed independent random variables is normal, we are also implicitly assuming that the first-innings scores are normally distributed. The true data generating process creates a score that is censored at zero. We note that assuming normality for performance and conditions raises the prospect of a negative total score; however, this is extremely unlikely over the range of the data.[4] The log-normal distribution, while having the desirable property of being bounded at zero, does not fit the data well. The assumed relationship between the mean and variance of our performance, conditions and first-innings score distributions is

$$\mu_S = \mu_\rho + \mu_\chi$$

$$\sigma_S^2 = \sigma_\rho^2 + \sigma_\chi^2$$

[4] The probability, given the mean and variance of our full data set, of our assumed normal distribution generating a score in any given match less than zero is 0.000016. This means that over our dataset of 784 matches, the probability of all our observed scores being greater than zero is 98.8%.

We centre the conditions variable around the mean first-innings score and the performance variable around zero in order to create the interpretation that a performance is a certain number of runs more or less than the conditions are worth. This approach, however, is simply a normalising assumption that we make without loss of generality.

4.2.1 Inferring the values of σ_ρ^2 and σ_χ^2

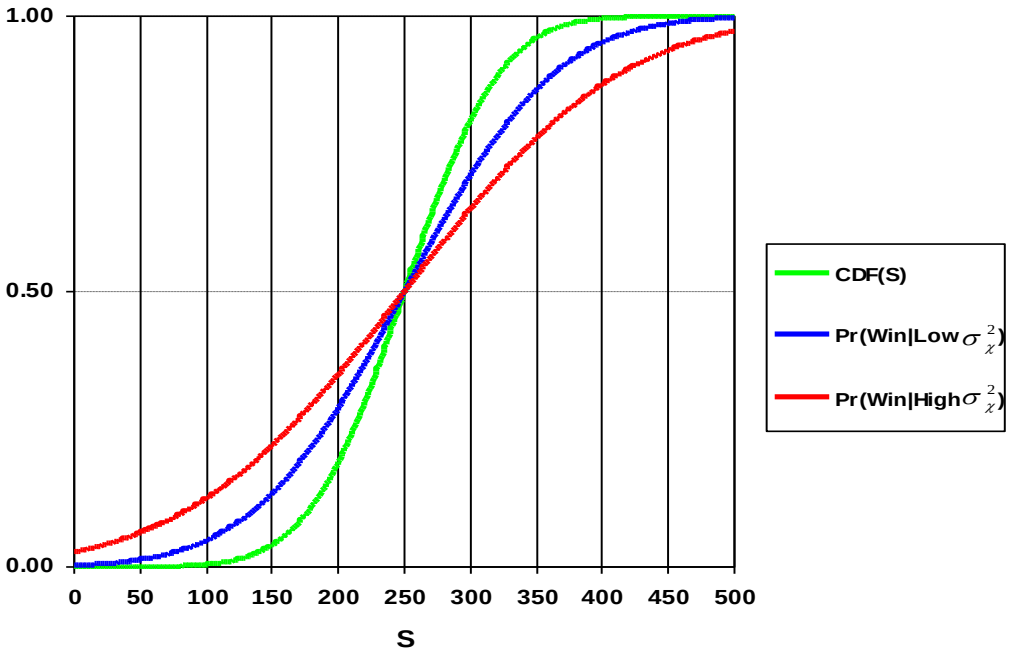
In order to show how we might go about estimating σ_ρ^2 and σ_χ^2, consider two hypothetical games. In hypothetical game one, we assume that conditions are known and do not change from match to match $(\sigma_\chi^2 = 0)$. Team 1 draws a number from the performance distribution ρ and Team 2 then draws a number from the same distribution. The team drawing the higher value of ρ wins and the first-innings score *S* is equal to $\chi + \rho$. In this game, the probability of Team 1 scoring fewer runs than a particular score is exactly the same as their probability of successfully defending that score in the second innings. That is, the graph of the cumulative distribution of first-innings scores will be identical to the graph showing the probability that a team with a given score in the first innings will win the game.

In hypothetical game two, we allow conditions to have a positive variance $(\sigma_\chi^2 > 0)$. This time nature draws a value for χ before the game begins and the two teams subsequently draw values for ρ. As with hypothetical game one, the team drawing the higher value of ρ wins and the first-innings score *S* is equal to $\chi + \rho$. In this game, however, the presence of variability in conditions will affect both the observed distribution of *S* and the probability of each score being a winning one. If we only observe the distribution of first-innings scores, that is, we do not observe any information about performance or conditions, the scores achieved

contain information about the conditions. If we observe a lower than average first-innings score, it is more likely that this game was played under a relatively low draw from the conditions distribution, when compared to our prior of no knowledge about conditions. That is, conditions, more likely than not, were more difficult for batting than average and bear part of the responsibility for the lower than average score. When a higher than average first-innings score is observed, it is more likely than not a match played in better than average batting conditions and part of the responsibility for the high score belongs to the conditions, rather than being solely allocated to the performance of the two teams.

The conditions variance affects the second-innings probability of winning function. The observation of a low score increases the probability that this match is being played under a low draw from the conditions distribution, which means that the probability of Team 1 successfully defending the score is higher than the *a priori* probability of scoring fewer than that score in the first innings when nothing is known about the conditions. The probability of winning function is therefore flatter than the cumulative density of scores function where we have a non-zero variance of conditions. The higher is the variance of conditions, the flatter is the probability of winning function, assuming a constant total variance of scores. We illustrate this in Figure 4.1.

Figure 4.1: First-innings Score and Probability of Winning



At this point, we need to define the second-innings distribution as the function whose cumulative density function is identical to the probability of winning function.

$$J(S) = \Pr(\omega \mid S)$$

where

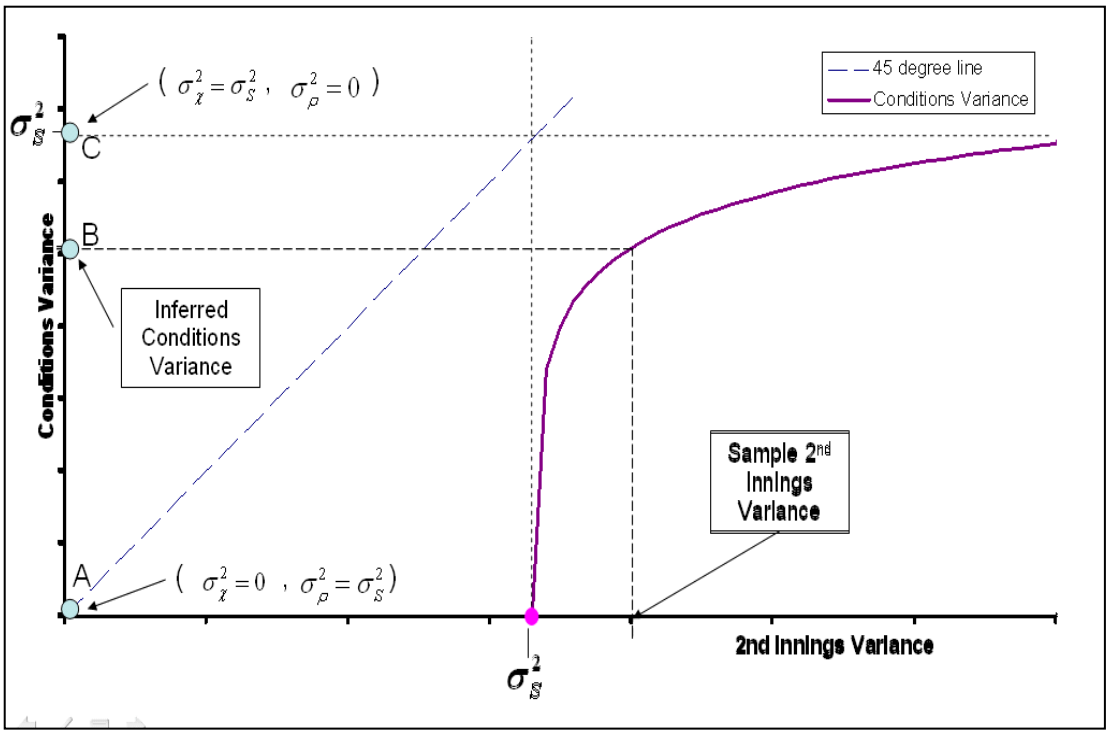
$$\omega = 1, \quad if\ Team1\ wins\ the\ match$$
$$\omega = 0, \quad if\ Team1\ loses\ the\ match$$

Note that this is a very different concept to the distribution of the actual scores observed in the second innings, which we do not use in this paper other than in determining the value of ω for each game. The actual second-innings scores are not of significant informational value as the game ends when Team 2 moves ahead of Team 1's score as further

play from that point is redundant given that Team 2 has already won the match. The distribution of Team 2's totals would simply be a lower bound to their potential scores in completed innings.

Once we have the first and second-innings distributions, we are able to infer the contributions of σ_χ^2 and σ_ρ^2 to σ_S^2 by comparing the variances of the first and second-innings distributions. We show this in Figure 4.2 where we plot combinations of conditions variance and second-innings variance for a given value of the first-innings score variance.

Figure 4.2: Inferring the contribution of conditions variance for a given σ_S^2



A conditions variance of zero (see Point "A") will lead to the second-innings variance being equal to the first-innings variance; this is the case in hypothetical game one, where the variation of performance explains the total variation of score. At the other extreme, a conditions

variance tending to the first-innings score variance (see Point "C") will lead to the second-innings variance tending to infinity. In this case, the entire variation in score is due to the conditions and the level of performance is constant. If we know the value of the second-innings variance then we can read the value of the implied conditions variance from the vertical axis of the graph (see Point "B").

4.2.2 Accounting for the second-innings advantage

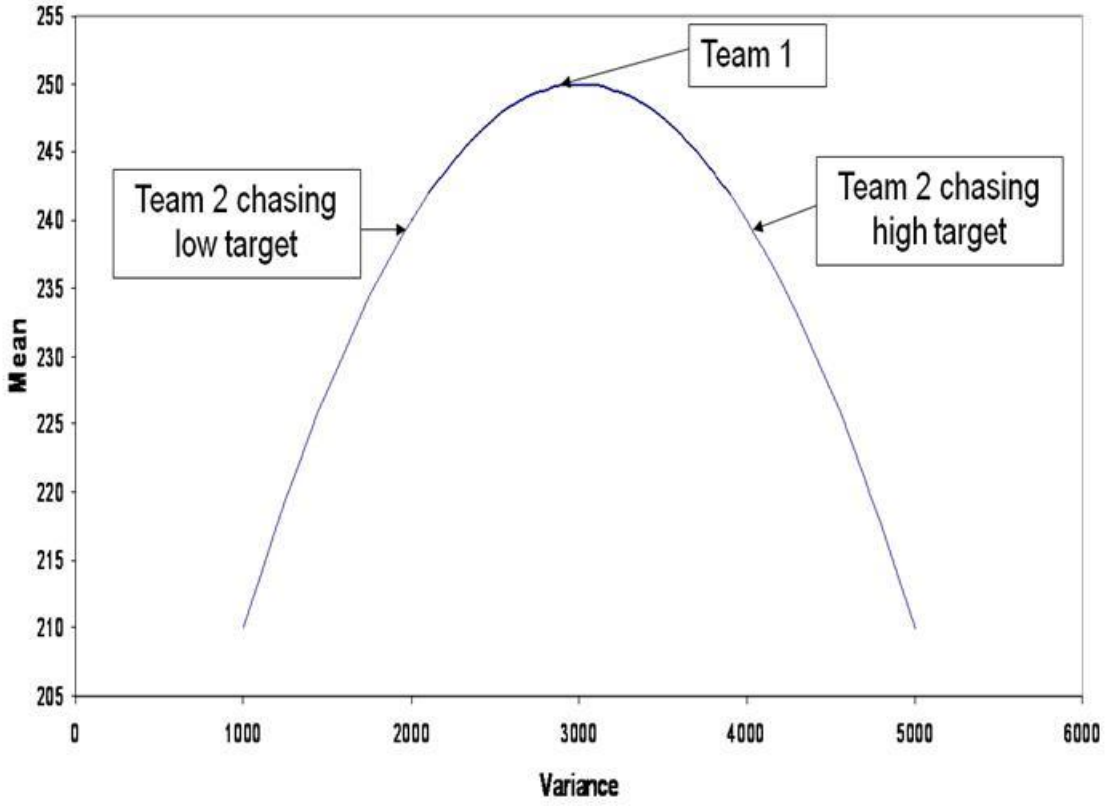
We need to make an additional adjustment before we can begin to estimate the values for σ_{χ}^2 and σ_{ρ}^2. Hypothetical game two assumes that the two teams draw from a distribution of ρ whose mean and variance are exogenous; that is, both teams are drawing from the same distribution so there is no advantage in the order of drawing. However, there is a theoretical second-mover advantage in ODI cricket. This is due to the team batting second having a known target score, resulting in them being able to adjust their risk strategy depending on the target. We acknowledge that while the fielding team do have some control over the overall risk strategy of the innings, in terms of bowling style and field placement, significantly more control over the risk strategy is available to the batting team. This is obvious to any cricket watcher as we almost always see the scoring rate increase and the survival rate decrease towards the end of the first-innings, which is what the team batting first would generally prefer as its overs begin to run out.

We can view the second-innings advantage as resulting from the batting team choosing a level of risk. The higher is the chosen level of risk, the higher is the scoring rate but the lower is the probability of survival. While the team batting first wishes to, in most situations,

maximise their expected total score and therefore chooses the level of risk resulting in the highest expected score, the team batting second wishes to maximise the probability of scoring a higher total than the team batting first achieved. The binary nature of the outcome of the second innings ensures that a team chasing a high total optimises by choosing a high level of risk, which is the equivalent of drawing from a distribution of performance with high variance, while a team chasing a low total optimises by choosing a low level of risk - that is, drawing from a low variance distribution. In both cases the optimal distribution from which the second mover draws ρ would likely have a lower mean than the optimal distribution from which the first mover draws ρ.

We outline this concept in Figure 4.3. A team chasing 200 runs in conditions which are worth 250, for example, might optimise by drawing from a distribution with a mean of 240 and a variance of 2000 runs, rather than the optimal first-innings strategy which might have a mean of 250 and a variance of 3000 runs. They accept a lower mean by taking a lower level of risk but in doing this they increase their probability of scoring 201 or greater. By contrast, a team chasing 300 runs in conditions which are worth 250 might optimise by drawing from a distribution with a mean of 240 and a variance of 4000 runs. Their high risk level reduces their expected score but the high variance increases the probability of getting the extremely large score that they require to win the match.

Figure 4.3: Potential choices of over risk level for an innings



We assume that the second-innings advantage is the difference between the means of the first and second-innings distributions and that it is a constant number of runs regardless of the first-innings score. If Team 1 scores an average score, then Team 2's optimal strategy may well be to bat as if they were batting in the first innings; that is, it would choose the level of risk that maximises the expected score. However, this assumes that Team 2 only gets to select their risk level once, at the beginning of their innings. This is not the case in a game of cricket; Team 2 can adjust their risk level at any time, depending on how their innings is progressing. For this reason, a constant second-innings advantage, independent of the first-innings score, is a reasonable assumption. We will incorporate this second-innings advantage into our $J(S)$ functions in our conditional probability formula for conditions given score and result.

4.3 The Data

4.3.1 Sources and timeframe

The research described in this chapter only requires three pieces of information: the date that the match was played; the first-innings score; and the result of the match. This information is publicly available on www.cricinfo.com. We select our time period as the decade of the 2000s: from January 1, 2000 until December 31, 2009. There were a total of 1405 official One Day Internationals played during this decade. Subsequent chapters of this thesis require more detailed ball-by-ball data and this is a subset of the data used in this chapter.

In order to ensure a robust analysis, there are some additional factors to consider when selecting the data set. As at the date of writing, there are sixteen countries with official ODI status.[5] It is generally accepted among cricket followers that there is a significant gap between the top-eight-ranked countries in the world and the remaining countries. We therefore only select matches played between two top-eight countries in our data set. Additionally, to perform the analysis we need an estimate of the distribution of first-innings scores in completed innings. On occasion, rain interferes in the game of cricket, resulting in a shortened match or even causing the complete abandonment of the match. These matches have the potential to distort our analysis. In order to be included in our data set, at match must meet all of the following criteria:

- the match was played between January 1, 2000 and December 31, 2009, inclusive;

[5] These teams are Australia, Afghanistan, Bangladesh, Canada, England, India, Ireland, Kenya, Netherlands, New Zealand, Pakistan, Scotland, South Africa, Sri Lanka, West Indies and Zimbabwe.

- the match was between two top-eight countries;
- the first innings was not shortened in any way other than the batting team being bowled out before their full allotment of 50 overs had been used; and
- the match was not abandoned without the declaration of a winner.

The total number of matches meeting all these criteria is 784. This forms our dataset for the analysis in this chapter.

4.3.2 Data demographics

We outline the number of matches involving each team and in each venue country in Tables 4.1 and 4.2. These tables show that we have a good distribution of matches.

Table 4.1: Number of matches played by each team

Country	Bat First	Bat Second	Total
Australia	130	98	228
England	90	78	168
India	106	123	229
New Zealand	82	103	185
Pakistan	103	104	207
South Africa	83	109	192
Sri Lanka	120	85	205
West Indies	70	84	154

Table 4.2: Number of matches played in each country

Country	Matches
Australia	122
England	81
India	99
New Zealand	73
Other	91
Pakistan	53
South Africa	110
Sri Lanka	83
West Indies	72

4.3.3 Two approaches

In the remainder of this chapter we show the results that we obtain from splitting the overall distribution of first-innings scores into their performance and conditions components. We show two possible approaches. Firstly, we simply perform the analysis using our entire data set. Secondly, we split the data into various segments of time in order to test for the impact of changes to the rules of the game.

4.4 Approach One: A single dataset

In this section we analyse our 784 matches as one single dataset, assuming that changes to the rules of ODI cricket over the ten-year period have no impact and that all eight teams are of equal ability. The advantage of making these assumptions is they allow us to work with the largest possible dataset with which to estimate our distributions.

4.4.1 Testing for normality of first-innings scores

Before we can attempt to split the variance of first-innings scores into their performance and conditions components we need to test our assumption that these scores are approximately normally distributed. Figure 4.4 shows the frequency of the first-innings scores, in bins of thirty runs, while Figure 4.5 compares the CDF of the first-innings score data with the CDF of a normal distribution with the same mean and variance. Normality appears to be a reasonable assumption; however, we perform a statistical test to confirm our conclusion.

Figure 4.4: Distribution of first-innings scores

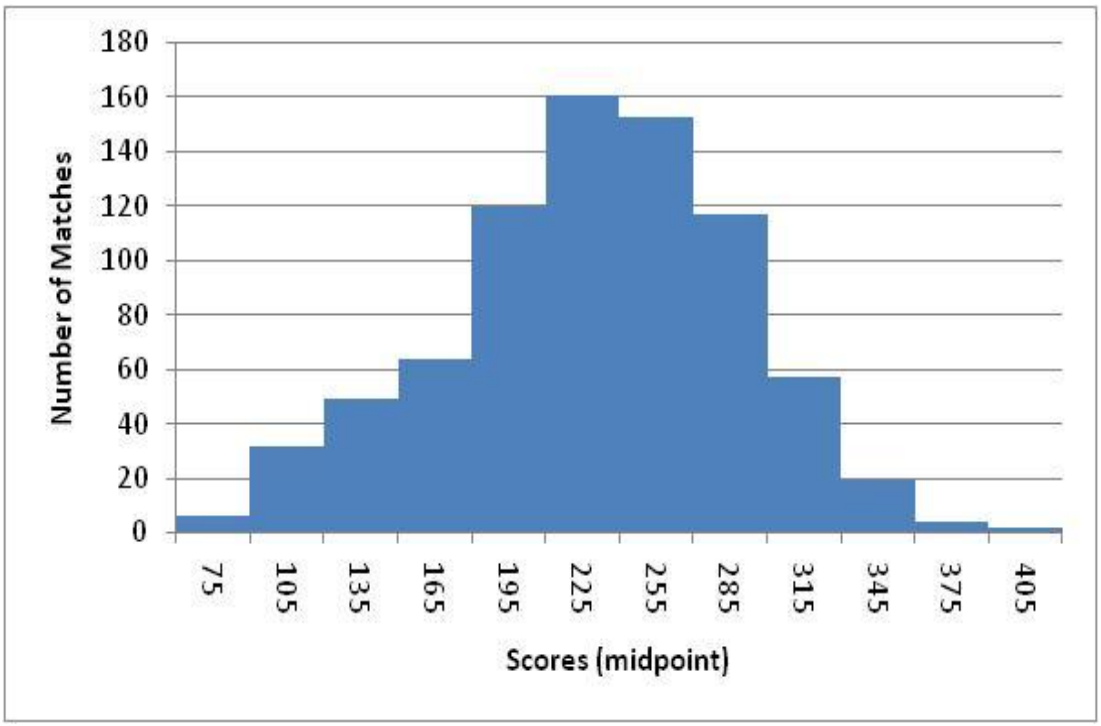
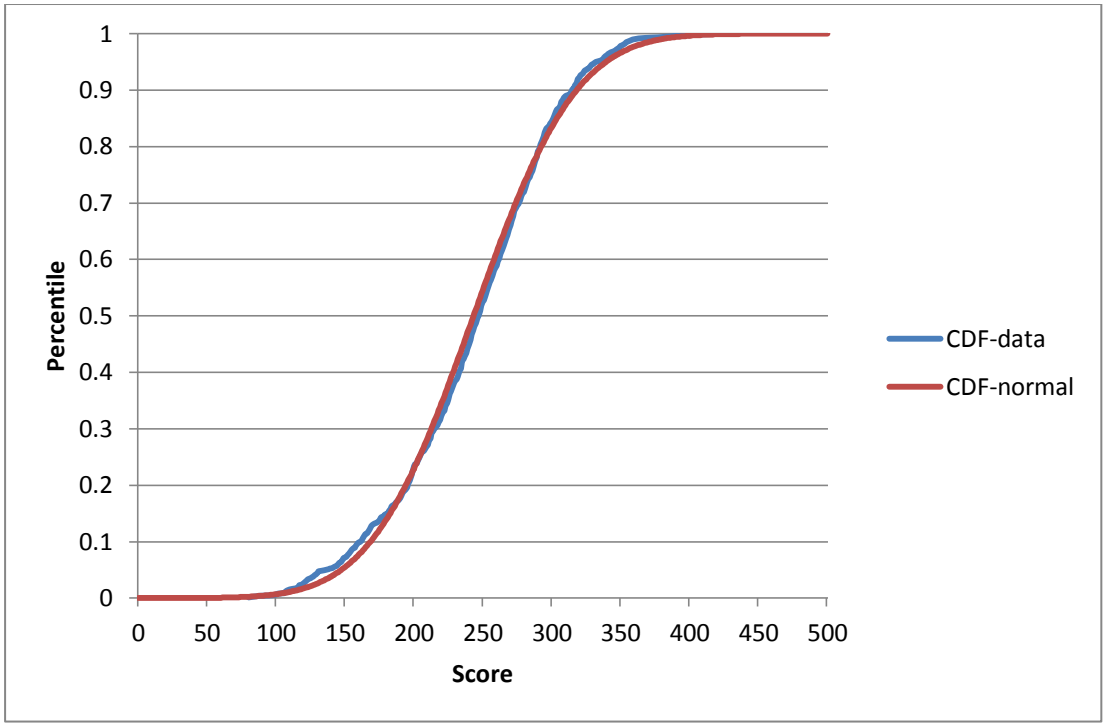


Figure 4.5: CDF comparison: Data vs. Normal Distribution



The Jarque-Bera test uses the sample skewness and kurtosis to test potential departures from normality. The null hypothesis is that the data are normally distributed, implying that both skewness and excess kurtosis are equal to zero. At this point it is useful to show some summary statistics of the data, in Table 4.3.

Table 4.3: Summary Statistics for First-innings Scores

Statistic	Value
n	784
Mean, (μ_S)	243.3
Median	247.5
Variance, (σ_S^2)	3412.5
Standard Deviation, (σ_S)	58.5
Skewness	-0.228
Kurtosis	2.888
Excess Kurtosis	-0.112002

The Jarque-Bera test statistic is

$$JB = \frac{n}{6}(\varsigma^2 + \frac{1}{4}K^2) \quad (1)$$

where n is the number of observations in the sample, ς is the sample skewness and K is the sample excess kurtosis. The Jarque-Bera statistic has an asymptotic chi-square distribution with two degrees of freedom. This chi-square distribution is an approximation of the true distribution of the Jarque-Bera statistic and is prone to making Type I errors. We identify the true distribution of the Jarque-Bera statistic for a sample size of 784 by Monte Carlo simulation. We generate 784 values from the standard normal distribution, calculate the skewness and excess kurtosis before finally calculating the JB statistic. Repeating this process 10000 times gives us a distribution of 10000 JB statistics under the assumption of normality. We are asking the question, were our data normal, how likely would we be to get a JB statistic as extreme as the one we observe by random chance alone in a sample of the same size as ours. In our data set, $\text{JB} = 7.186640$. This value occurs between the 9692$^{\text{nd}}$ and 9693$^{\text{rd}}$ observations of our simulated distribution of 10000 JB statistics and therefore we are able to reject the null hypothesis that the data are normally distributed at the 5% significance level.

It is important to see how much our scores deviate from the normal distribution with the same mean and variance. Sorting the data in ascending order of first-innings score, if the data are normally distributed the n^{th} score should be equal to the inverse normal of $\frac{1}{n}$, for our mean and variance. Ignoring the first and last five observations in a bid to eliminate any outliers, the mean absolute deviation of our observed score from the theoretical score implied by the normal

distribution is 4.1 runs. This indicates that our data are, on average, not substantially different to the normal distribution.

The goal of this analysis is to estimate a value for conditions where no variable exists in the data set. We are not looking for perfection; rather, we are seeking a substantial improvement on the alternative of simply admitting to having no knowledge about conditions. While we have rejected the null hypothesis of perfect normality, it is clear from Figures 4.4 and 4.5 that the normal distribution is the best approximation to the data that we have. For these reasons, we proceed with our analysis despite the imperfect normality.

4.4.2 Estimating the second-innings variance

We have, from our 784 matches, a distribution of first-innings scores as well as the match result for each one of those 784 scores. We want to estimate the probability of winning for a given first-innings score. In order to achieve this, we construct a Probit model where we regress the outcome of the game on the score achieved by the team batting first. We have a very small sample (six) of tied matches; therefore we do not want to run an ordered Probit model with three outcomes as we would not get a good estimate of the relationship between first-innings score and the probability of a tie. We are not particularly interested in this outcome in any case. In order to simplify the analysis we repeat each tied match in the data set as one win and one loss and give each of these observations a weight of 0.5. All other observations have a weight of one in the regression, meaning that each match has a total weight of one. We define the probability of winning given first-innings score S function $\Pr(\omega \mid S)$ as a simple Probit model in Equation (2). Let

$$\omega = \begin{cases} 1 & \text{if the team batting first wins the game} \\ 0 & \text{if the team batting second wins the game} \end{cases}$$

$$\Pr(\omega = 1 | S = S_i) = \Phi(\alpha + \beta S) \tag{2}$$

where Φ is the cumulative distribution function of the standard normal distribution. It follows that

$$\Pr(\omega = 0 | S = S_i) = 1 - \Pr(\omega = 1 | S = S_i)$$

In our data set, $\alpha = -3.292363342$ and $\beta = 0.0132766688$, which means our Probit model is

$$\Pr(\omega = 1 | S = S_i) = \Phi(-3.292363342 + 0.0132766688 S)$$

This Probit model reveals the probability of winning for the team batting first, given that they scored a particular total. These probabilities given an implied distribution of second-innings scores, hereafter referred to as the second-innings distribution. To estimate the mean and variance of this second-innings distribution, μ_{S_2} *and* $\sigma^2_{S_2}$,we return to our Probit function, $\Pr(\omega = 1 | S = S_i) = \Phi(\alpha + \beta S)$, which implies the Z-score, *Z,* is calculated as

$$Z = \frac{S - \mu_{S_2}}{\sigma_{S_2}}$$

$$Z = \alpha + \beta S$$

In a standard normal distribution, Z is equal to zero at the mean value of S; therefore

$$0 = \alpha + \beta\mu_{S_2}$$

$$\Rightarrow \quad \mu_{S_2} = \frac{-\alpha}{\beta}$$

We substitute the mean into the formula for Z in order to calculate the variance as

$$\frac{S + \frac{\alpha}{\beta}}{\sigma_{S_2}} = \alpha + \beta S$$

$$\Rightarrow \quad \frac{\beta S + \alpha}{\beta\sigma_{S_2}} = \alpha + \beta S$$

$$\Rightarrow \quad \sigma_{S_2}\beta = 1$$

$$\Rightarrow \quad \sigma_{S_2} = \frac{1}{\beta}$$

$$\Rightarrow \quad \sigma_{S_2}^2 = \frac{1}{\beta^2}$$

We are now able to plug in our values for α and β to determine the mean and variance of our second-innings distribution.

$$\mu_{S_2} = \frac{-\alpha}{\beta} = 247.981$$

$$\sigma_{S_2}^2 = \frac{1}{\beta^2} = 5673.117$$

4.4.3 Splitting the first-innings variance

A summary of the information required to split the first-innings variance into its separate performance and conditions components is provided in Table 4.4. Note that there is a difference between the first-innings mean and the second-innings mean of 4.7 runs. Also, note that the second-innings variance is significantly higher than the first-innings variance. These differences will play a major role in determining our measures of conditions and performance.

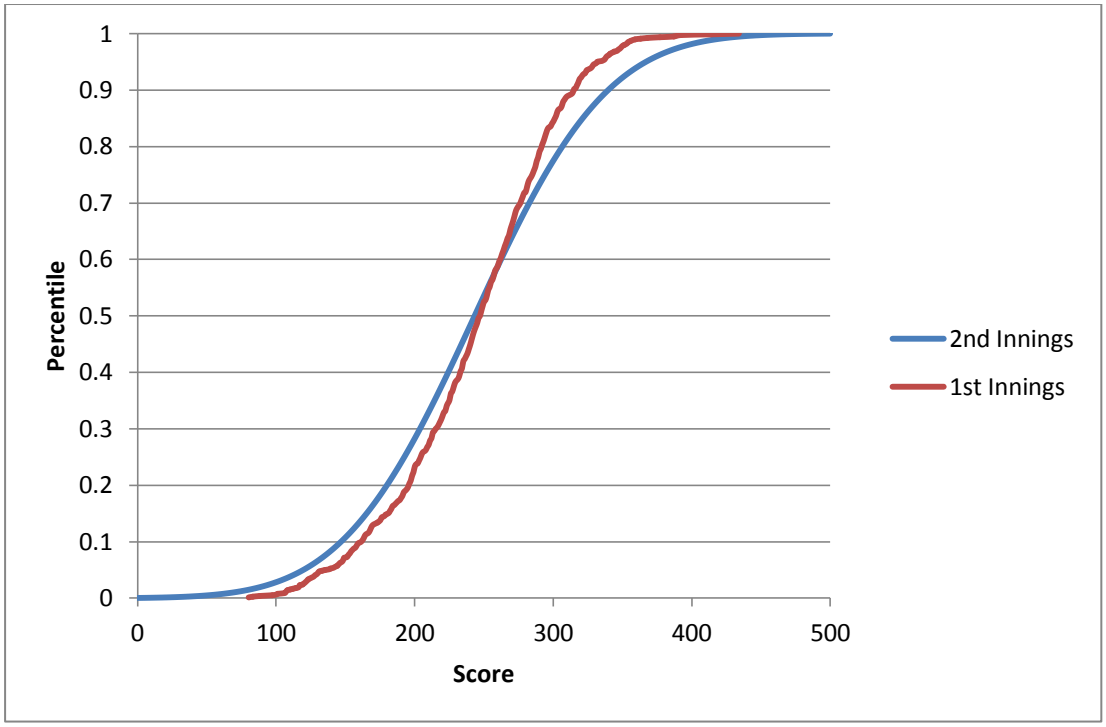
Table 4.4: Means and Variances of the distributions

	N	Mean 1st Innings	Variance 1st Innings	Mean 2nd Innings	Variance 2nd Innings
Score	784	243.287	3412.488	247.981	5673.117

Figure 4.6 shows the differences between the first and second-innings distributions. We have linearly adjusted the second-innings distribution to remove the second-innings advantage of 4.694 runs. It is clear that the cumulative distribution functions cross over at approximately the 50% mark.[6] The implication of this crossover is that, after the removal of the second-innings advantage, scores in the upper ranges of the distributions are easier to chase successfully than they are to score in the first innings, where nothing is known about conditions prior to the first-innings score being observed. Conversely, scores in the lower ranges of the distributions are more difficult to chase successfully than they are to score. This is due to the second-innings distribution having a greater variance.

[6] The crossover point is not exactly at the 50th percentile due to the small amounts of non-normality present in the first-innings distribution. The second-innings distribution, being sourced from a Probit model, is normal by definition.

Figure 4.6: Adjusted cumulative distributions



To split the first-innings variance (σ^2_S = 3412.488) into performance variance (σ^2_ρ) and conditions variance (σ^2_χ), we investigate combinations of σ^2_ρ and σ^2_χ in order to find a combination that would result in a second-innings distribution with similar variance to that implied by our Probit model. This enables us to construct a distribution of first-innings score S as a function only of performance ρ, where this new distribution represents the level of performance that is required on average to achieve each score. As this distribution assumes that conditions are unknown, this is exactly the situation that we are faced with when we estimate the second-innings distribution. It follows that this distribution should approximate the second-innings distribution; since we are assuming that the conditions are the same for both teams then if ρ is higher in the second innings then the team batting second should win. Note that in this analysis we ignore the second-innings advantage. Later in this chapter we show that a constant value second-innings advantage has no impact on the second-innings variance.

For each possible share of the total variance that we investigate, we are asking the question "if these were the shares of the total variance attributable to each factor, what would the second-innings variance be?" We then select the values of σ_ρ^2 and σ_χ^2 that result in the closest second-innings variance to that observed in our data set.

To determine the second-innings variance that would exist under each possible split of the first-innings variance, we need to define some variables. Let ω be a binary variable taking a value of one if the team batting first wins the match and zero otherwise. Let $f(\chi)$, $g(S)$ and $k(\rho)$ denote the density functions of χ, S and ρ respectively, and let $K(\rho)$ denote the distribution function of ρ. Let $\Pr(\omega)$ be the probability of observing outcome ω. Further, let $\hat{f}(\chi \mid S)$ be the conditional density of χ given S. Let δ be the fraction of the first-innings variance allocated to performance, which implies that a fraction of $(1-\delta)$ of the first-innings variance is allocated to conditions and let $J_\delta(S)$ be the distribution function of S implied by the probability of winning function $\Pr(\omega = 1 \mid S)$ under the assumption of a split of δ. Finally let S_2 be a latent variable which is a measure of performance and is the sum of ρ_2 and χ. This latent variable indicates a win for Team 1 if $S_2 < S$ and has the property $J_\delta(S_2) = J_\delta(S)$.

The probability that Team 1 wins, having scored S, is equal to the total of the probabilities of Team 1 winning having scored S in conditions worth χ, over all possible values of χ; that is

$$J_\delta(S) = \Pr(\omega = 1 \mid S) = \int \Pr(\omega = 1 \mid S, \chi) \cdot \hat{f}(\chi \mid S) d\chi . \qquad (3)$$

Now, the probability of Team 1 winning given they made a particular score, *S*, in particular conditions χ, is equal to the probability that Team 2 achieves a lower ρ than that achieved by Team 1. Since $S = \rho + \chi$, we can write

$$\Pr(\omega = 1 | S, \chi) = K(S - \chi) \tag{4}$$

Given that Team 1 scores *S*, the probability that they were playing in conditions worth χ is equal to the density of χ multiplied by the density of performance $\rho = S - \chi$, divided by the overall density of *S*.

$$\hat{f}(\chi | S) = \frac{f(\chi) \cdot k(S - \chi)}{g(S)} \tag{5}$$

When we put Equations (3), (4) and (5) together we get

$$J_{\delta}(S) = \int K(S - \chi) \cdot \frac{f(\chi) \cdot f(S - \chi)}{g(S)} d\chi. \tag{6}$$

We perform some numerical investigations into the properties of Equation (6), included as Appendix A. It turns out that the distribution defined by Equation (6) is normal, with variance defined by

$$\sigma^2_{S_2,\delta} = \sigma^2_S \left(\frac{2}{\delta} - 1 \right)$$

Given that we know $\sigma^2_{S_2}$ and σ^2_S, we can define the fraction of score variance that should be allocated to performance as

$$\delta = \frac{2\sigma_S^2}{\sigma_S^2 + \sigma_{S_2}^2} \tag{7}$$

Equation (7) enables us to determine the appropriate split of the variance of the first-innings scores into its performance and conditions components for any first-innings variance and Probit-implied second-innings variance.

Table 4.5 outlines the variances of performance and conditions and their shares of the total first-innings variance.

Table 4.5 Estimated split of the first-innings variance

σ_S^2	σ_ρ^2	σ_χ^2	δ	1-δ
3412.5	2563.4	849.1	0.751	0.249

Now that we have obtained the variances of ρ and χ, we are able to combine this information with the assumed means of zero and 243.3 (respectively) and assumed functional form of normality, in order to plot the distributions in Figure 4.7. Note that changing the assumptions that the performance mean is zero and the conditions mean is 243.3 would simply shift the performance and conditions distributions along the x-axis. As we have defined them, our functions are

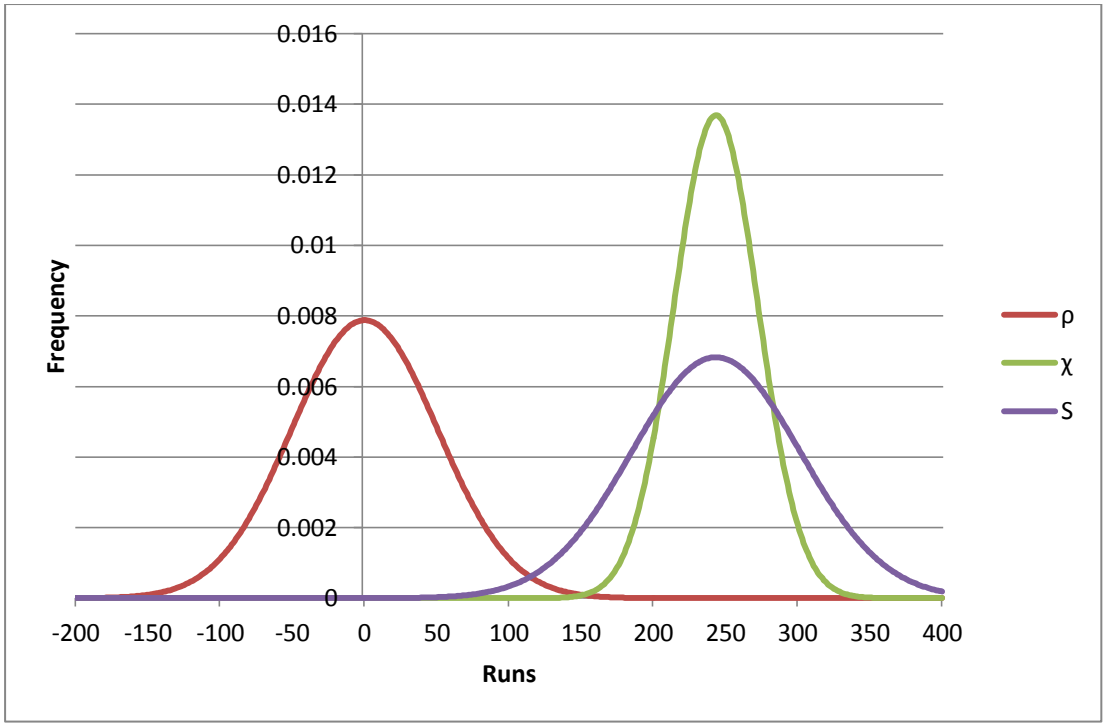
$$g(S) \sim N(243.287, 3412.488) \tag{8}$$

$$f(\chi) \sim N(243.287, 849.076) \tag{9}$$

$$k(\rho) \sim N(0, 2563.412) \tag{10}$$

$$J(S) \sim N(243.287, 5673.117)$$

Figure 4.7: The performance, conditions and score distributions



4.4.4 Determining the second-innings performance advantage

Previously we have modeled the probability of Team 1 winning given that they have scored *S* runs in the first innings as

$$J(S) = \int K(S-\chi) \cdot \frac{f(\chi) \cdot f(S-\chi)}{g(S)} d\chi. \tag{11}$$

In the presence of a second-innings performance advantage of a constant number of runs regardless of first-innings score, which we call A_ρ, Equation (11) changes to the following

$$J(S) = \int K(S - \chi - A_\rho) \cdot \frac{f(\chi) \cdot f(S - \chi)}{g(S)} d\chi. \quad (12)$$

Equation (12) incorporates the notion that the probability that Team 2 will lose when chasing S_1 runs to win is equal to the probability that Team 1 has of scoring fewer than $S_1 + A_p$ runs. That is

$$K(\rho_1) = K(\rho_2 - A_\rho)$$

Recall that the mean of the distribution of first-innings scores, μ_S, is equal to 243.3, while the mean of the implied second-innings distribution, μ_{S_2}, is equal to 248.0. The second-innings advantage observed in the number of runs scored is equal to the difference of these two means, 4.7 runs. We treat this as a constant rightward shift of the $J(S)$ function (which we previously modeled without the second-innings advantage), as shown in Figure 4.8. It is somewhat intuitive that this constant rightward shift could be achieved by an identical constant rightward shift of the performance distribution when modeling the batting innings of Team 2; however, this is not the case. To demonstrate, we apply the second-innings advantage to Equation (12) using our functions for performance and conditions obtained in the previous section.

$$J(S_{A=4.694}) = \int K(S - \chi - 4.694) \cdot \frac{f(\chi) \cdot f(S - \chi)}{g(S)} d\chi.$$

A magnified display of the outcome is shown in Figure 4.9. The $J(S)$ implied by the setting of the performance advantage A_ρ equal to the difference in the means of the first and second-innings distributions is to the right of the observed $J(S)$ distribution. This means that when Team 2 has an advantage in performance of a particular number of runs, A_ρ, the advantage observed in the distribution of scores implied by the probability of winning function is $A_S > A_\rho$. In our example, using $A_\rho = 4.7$ leads to a second-innings mean of 249.5 and a second-innings variance of 5673.1. The inclusion of a constant second-innings performance advantage has had no impact on the second-innings variance.

Figure 4.8: The second-innings advantage in score

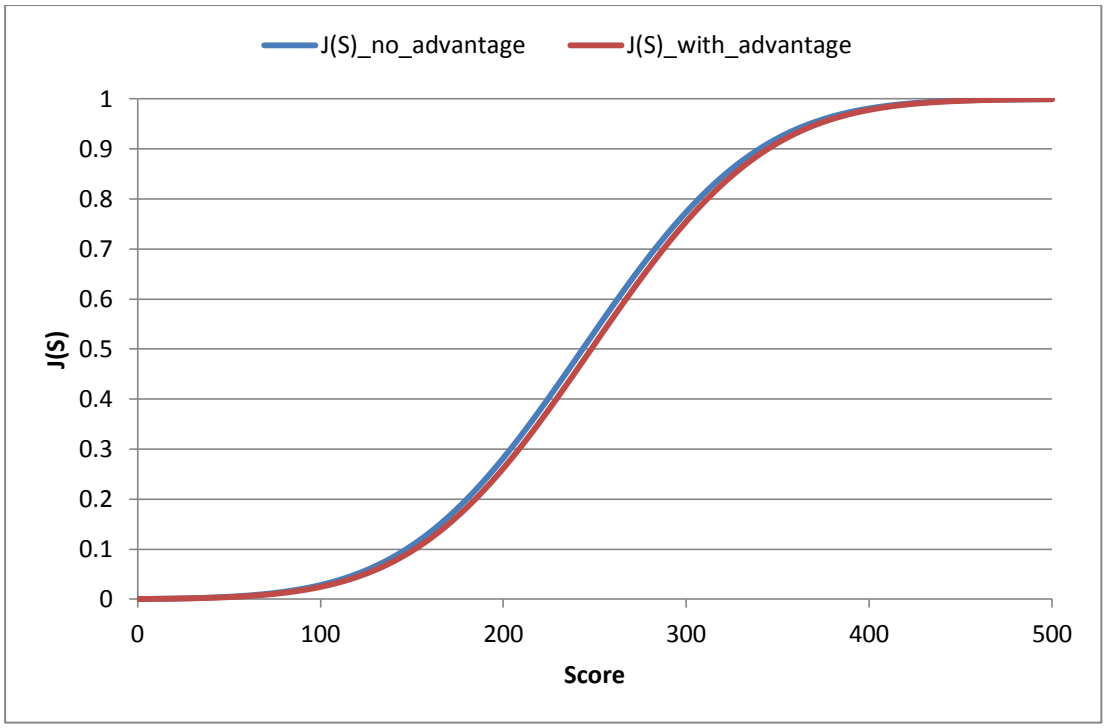
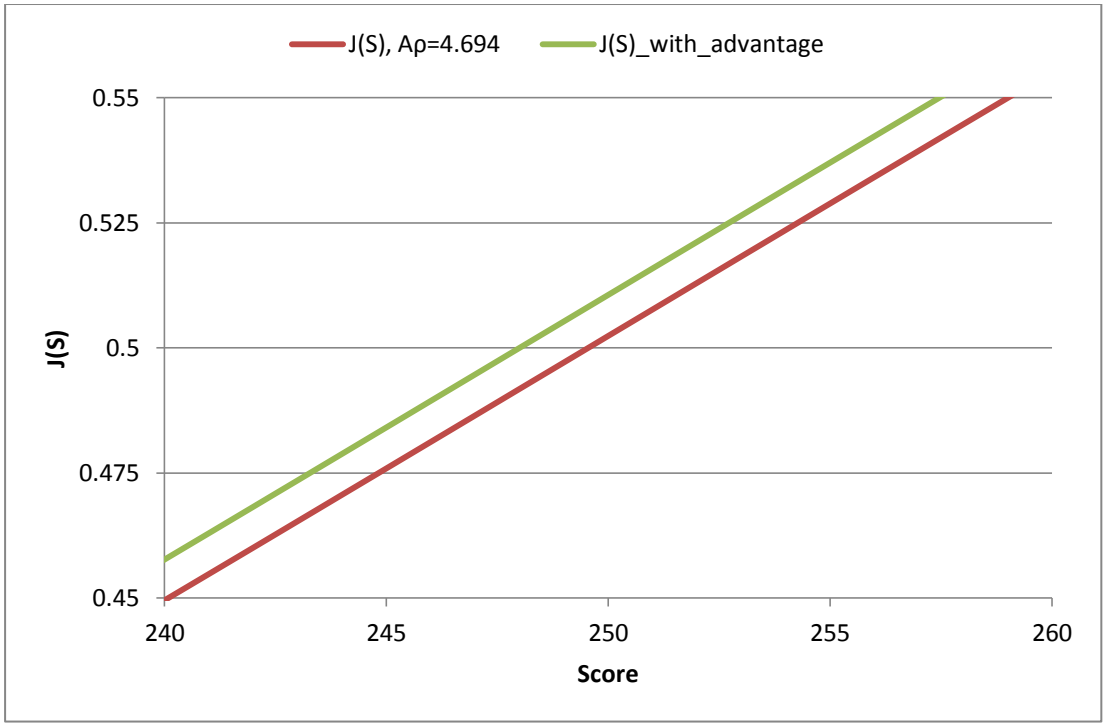


Figure 4.9: Imposing a performance advantage equal to the score advantage



In Appendix B we show some numerical simulations that are useful for determining the relationship between the performance advantage and the score advantage. It turns out that the mean of the second-innings distribution can be defined as

$$\mu_{S_2} = \mu_S + \frac{A_\rho}{\delta} \tag{13}$$

Equation (13) implies that the higher the fraction of first-innings score variance allocated to conditions, rather than performance, the larger the difference in the means of the first and second-innings distributions, for a given performance advantage. In our data set, $A_\rho = 3.526$.

4.4.5 Establishing the conditional distributions using Bayes' Theorem

We now have all the information that we need to determine the conditional distributions for conditions, given the first-innings score and the result of the game. Let $h(S,\omega)$ be the joint density function of S and ω. Using Bayes' theorem, we define the conditional density function for conditions in match *i* as

$$f(\chi = \chi_i \mid S = S_i, \omega = \omega_i) = \frac{f(\chi_i)g(S_i \mid \chi_i)\Pr(\omega_i \mid S_i, \chi_i)}{h(S_i, \omega_i)} \tag{14}$$

Since the probability of scoring *S* runs in conditions worth χ, $g(S \mid \chi)$, is equivalent to the probability of achieving performance $\rho = S - \chi$, Equation (12) can now be written as

$$f(\chi = \chi_i \mid S = S_i, \omega = \omega_i) = \frac{f(\chi_i)k(S_i - \chi_i)\Pr(\omega_i \mid S_i, \chi_i)}{h(S_i, \omega_i)} \tag{15}$$

Reproducing Equations (8), (9) and (10) from above and adding our calculated second-innings advantage, the equations that define the parameters of our model are

$$g(S) \sim N(243.287, 3412.488) \tag{8}$$

$$f(\chi) \sim N(243.287, 849.076) \tag{9}$$

$$k(\rho) \sim N(0, 2563.412) \tag{10}$$

$$A_\rho = 3.526 \tag{16}$$

We can estimate $f(\chi_i)$ and $k(S_i - \chi_i)$ directly from the distributions outlined in Equations (9) and (10). Since the probability of winning a match when batting first is equal to the probability of performing better than the other team by a greater amount than the second-innings advantage A_ρ, we can define the cumulative density function of performance as $K(\rho)$ and write

$$\Pr(\omega_i \mid S_i, \chi_i) = \begin{cases} 1 - K(S_i - \chi_i - A_\rho), & \omega = 0 \\ K(S_i - \chi_i - A_\rho), & \omega = 1 \end{cases}$$

Finally, $h(S_i, \omega_i)$ is the summation of the probability of observing each possible combination of performance, conditions and result over the set of outcomes where $S = S_i$ and $\omega = \omega_i$. This is equivalent to summing the numerator of Equation (15) over all outcomes where $S = S_i$ and $\omega = \omega_i$. We can write

$$h(S_i, \omega_i) = g(S_i) \Pr(\omega_i \mid S_i) \tag{17}$$

Since the probability of Team 1 winning with a score of *S* is *J(S)*, Equation (17) becomes

$$h(S_i, \omega_i) = \begin{cases} g(S_i)(1 - J(S_i)), & \omega = 0 \\ g(S_i) J(S_i), & \omega = 1 \end{cases}$$

This means that our final equations for determining the conditional density of conditions given the score and result are as follows

$$f(\chi = \chi_i \mid S = S_i, \omega = \omega_i) = \begin{Bmatrix} \dfrac{f(\chi_i)k(S_i - \chi_i)(1 - K(S_i - \chi_{i-} A_\rho))}{g(S_i)(1 - J(S_i))}, & \omega = 0 \\ \dfrac{f(\chi_i)k(S_i - \chi_i)K(S_i - \chi_{i-} A_\rho)}{g(S_i)J(S_i)}, & \omega = 1 \end{Bmatrix} \quad (18)$$

In words, Equation (18) is calculating the probability of achieving each particular combination of conditions, first-innings score and result, divided by the total probability of observing each combination of first-innings score and result. We use Equation (18) to determine the probability of a certain value of conditions given the first-innings score and the outcome of the game, two variables that are observable in the data set. Given a score and a result, we can determine $f(\chi \mid S, \omega)$ for every possible integer value of conditions and we have a random variable of which we can calculate the mean and variance using expected values.

4.4.6 Selected results

We plot selected examples of the conditional distributions of conditions given score and result in the following figures. The impact of the result of the game is shown in Figure 4.10, where the conditional distributions of conditions implied by two situations are shown, along with the naive prior distribution of conditions. In each situation, the team batting first scored 243 runs (the closest whole number to the overall mean in the data set), for one win and one loss. There are two important things to note about these distributions. First, the conditional distributions provide more certainty about what the conditions are like in each game, as their variances are substantially lower than the prior distribution. Second, knowing the result of the game makes a substantial difference to the mean of the conditional distribution, as outlined in Table 4.6. An average score resulting in a win shifts the conditional mean further from the prior

mean than an average score resulting in a loss, as there is a smaller than 50% chance of an average score resulting in a win, due to the second-innings performance advantage.

Figure 4.10: Inferred conditions under different match results

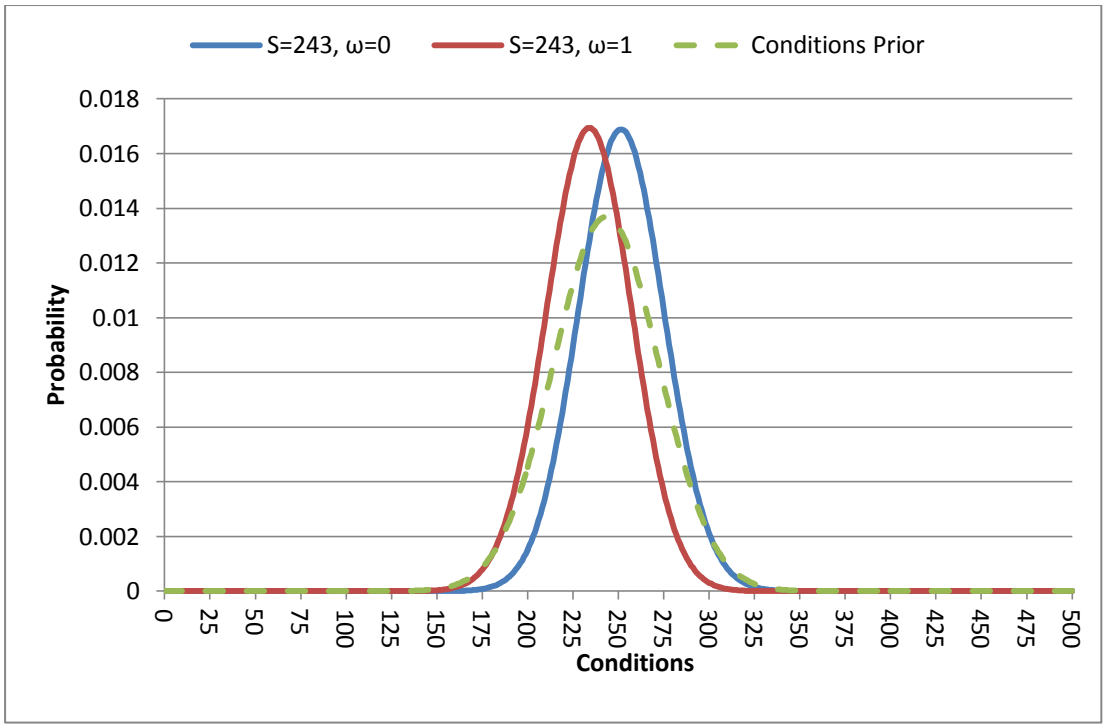


Table 4.6: Mean and Variance of inferred conditions under different match results

Conditions Distribution	Mean	Variance
Naïve Prior	243.3	849.1
$S = 243, \omega = 0$	251.7	558.8
$S = 243, \omega = 1$	233.7	555.1

In Figure 4.11 and Table 4.7, we show the impact of a particularly low first-innings score of 200 and a particularly high score of 300, both of which resulted in losses for Team 1. The conditional mean shifts much further away from the prior mean when 300 was scored as

for Team 1 to lose when they have scored a very high score is a surprising result. The variance is also lower in this situation, implying a greater level of certainty about the conditions.

Figure 4.11: Inferred conditions under different first-innings scores

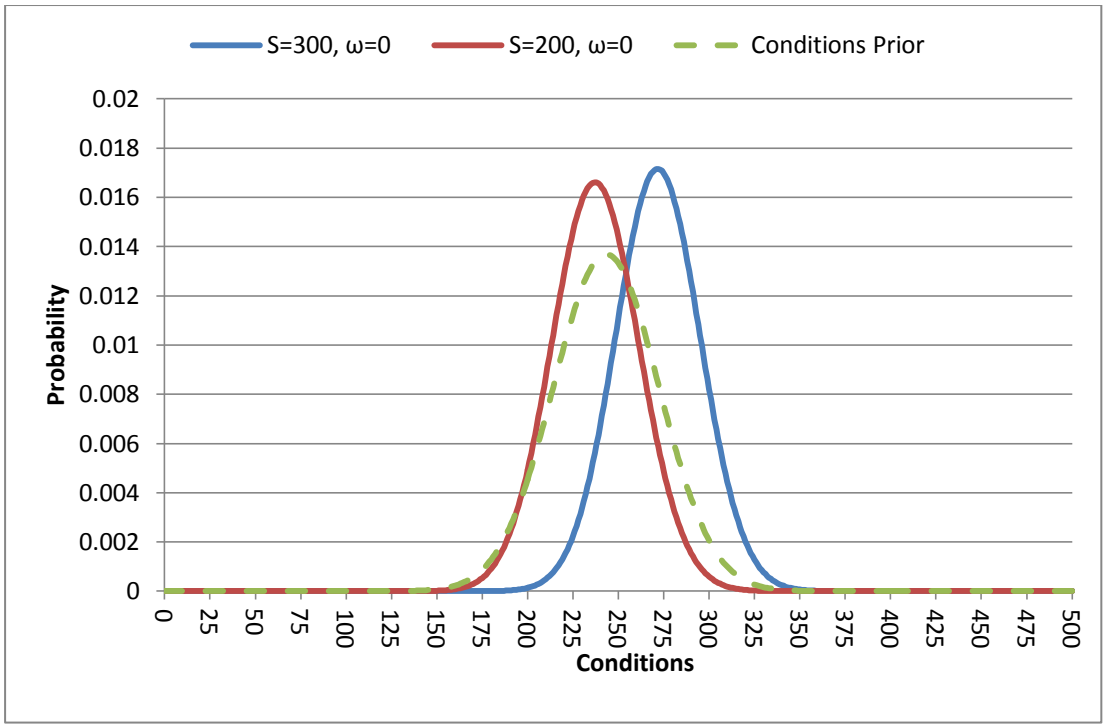


Table 4.7: Mean and Variance of inferred conditions under different scores

Conditions Distribution	Mean	Variance
Naïve Prior	243.3	849.1
$S = 200, \omega = 0$	237.5	577.3
$S = 300, \omega = 0$	271.9	541.1

More generally, we plot the means and variances of the inferred conditions distributions for each score and result of the game in Figures 4.12 and 4.13, respectively. As expected, the mean of the conditions distribution is higher in games lost by Team 1 than in games won by

Team 1, for a given first-innings total. We also note that the further away from the overall mean the first-innings score is, the larger the impact of one result compared with the other on the conditions distributions. Figure 4.15 shows that we have a higher level of certainty about the value of conditions when the result observed is the less likely one, given the first-innings score.

Figure 4.12: Inferred conditions means

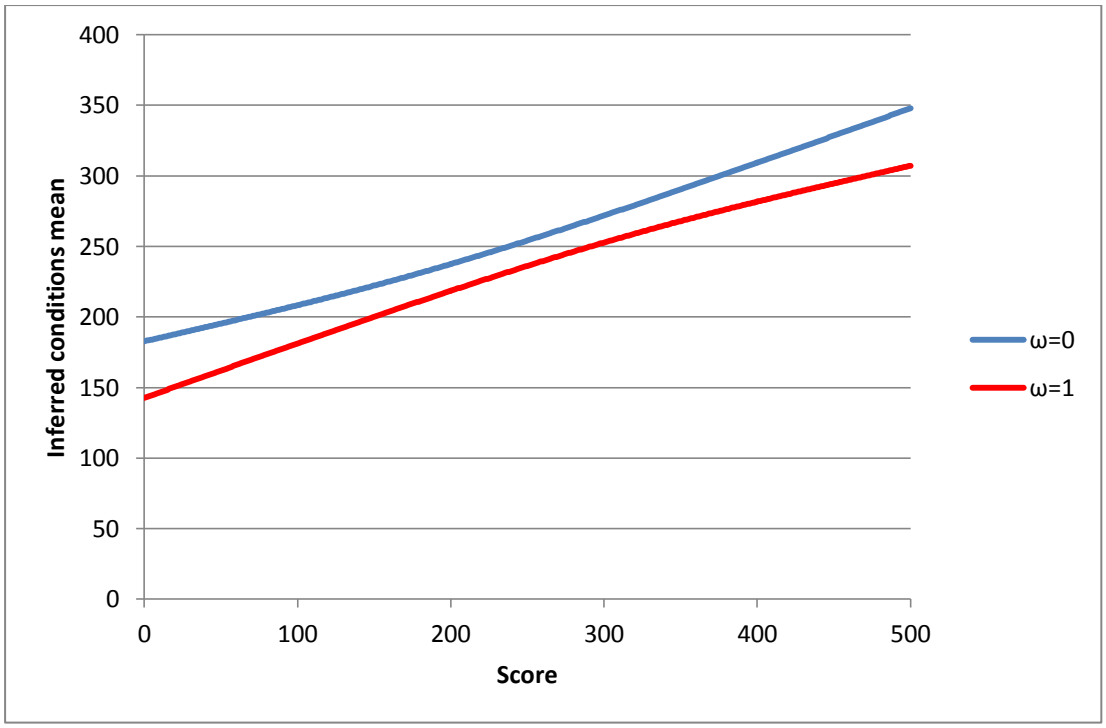
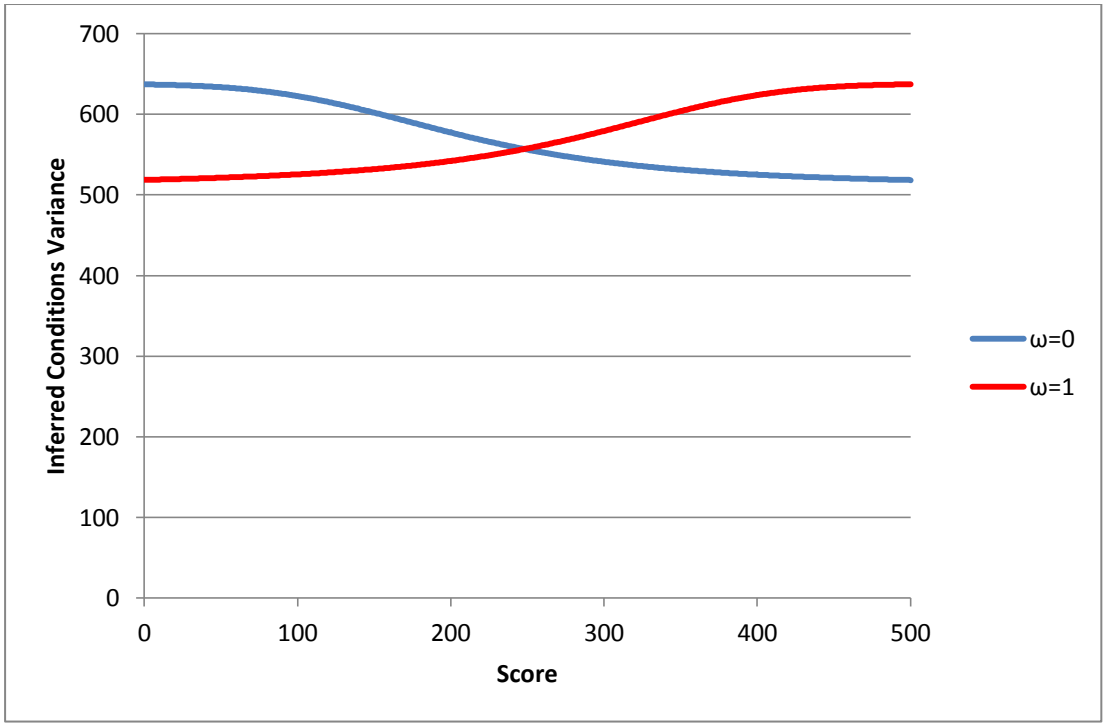


Figure 4.13: Inferred conditions variances



4.4.7 Testing the conditional distributions for normality

In subsequent chapters we are going to use our inferred conditions distributions in further models of the game. Since we have distributions, rather than a point estimate, we are going to make random draws from these conditional distributions, expanding the data set. It is useful to determine whether the conditional distributions are normal because if that is the case then we can make these random draws while needing only to store the mean and variance. Given our limited computing power, this would enable us to make substantially more draws than if we have to store the entire distribution numerically and compare each random draw against these numerical values.

We choose three situations from our analysis in the previous section to examine for normality. If the distributions are perfectly normal then they should have skewness and excess kurtosis equal to zero. Additionally, if we take Z-scores of the cumulative probability at each value of conditions, these Z-scores should be perfectly linear and therefore a linear regression through these Z-scores should have an R-square value equal to one. We show this information in Table 4.8.

Table 4.8: Normality checks for selected conditions distributions

Conditions Distribution	Skewness	Kurtosis	R^2 of Z-score OLS
$S = 200, \omega = 0$	0.0291	0.0034	0.999899
$S = 243, \omega = 1$	-0.0230	0.0060	0.999929
$S = 300, \omega = 0$	0.0164	0.0052	0.999962

Table 4.8 shows that for our three selected situations, the conditional distributions are very slightly skewed, but are hardly discernible from a normal distribution, with the R-square of the OLS regression of Z-scores on score being so close to one. More generally, we plot the skewness and excess kurtosis for all scores from zero to 500, in Figure 4.14 and Figure 4.15, respectively. We see that the conditions distributions are positively skewed when Team 1 loses and negatively skewed when Team 1 wins, with the skewness distributions themselves having the opposite skewness. The kurtosis distributions are more complicated; however, we see that regardless of the game result the excess kurtosis tends to be positive in the scores around the overall mean score of 243.3, where most scores would actually occur.

Figure 4.14: Skewness of conditions distributions

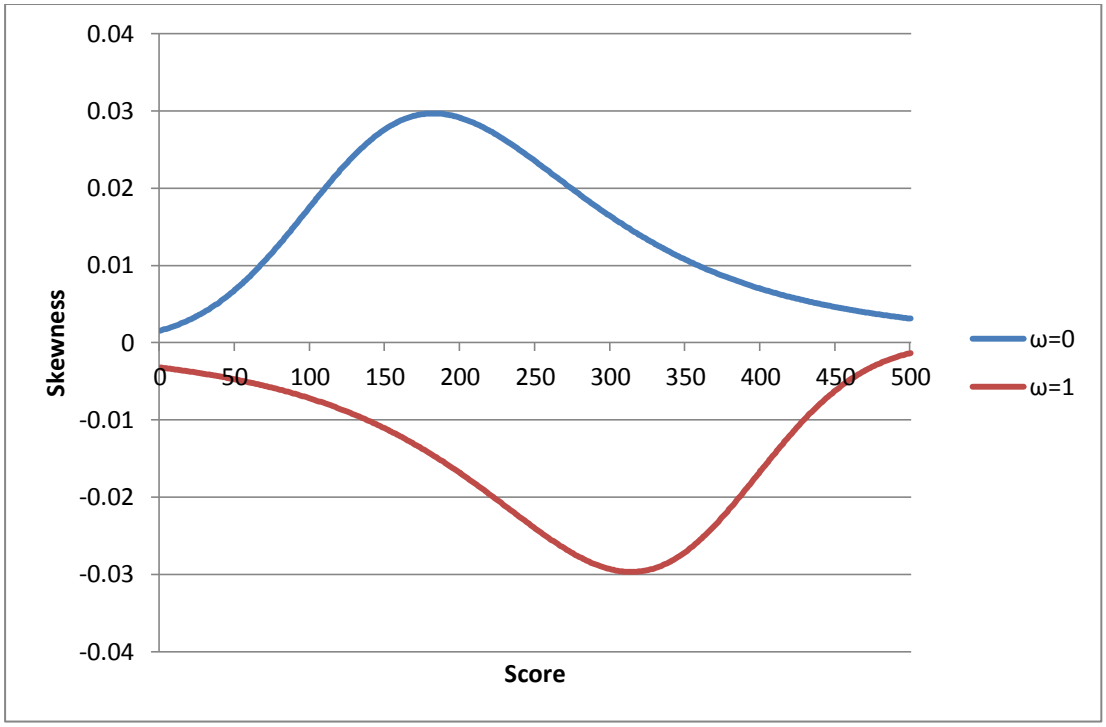
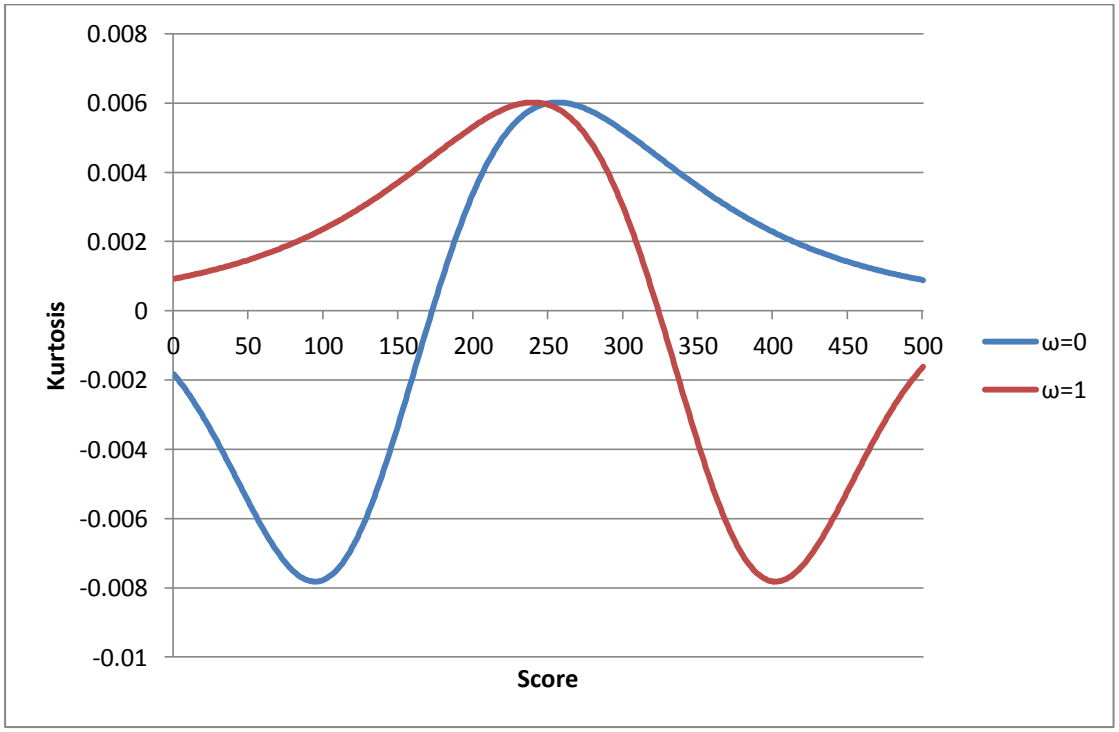
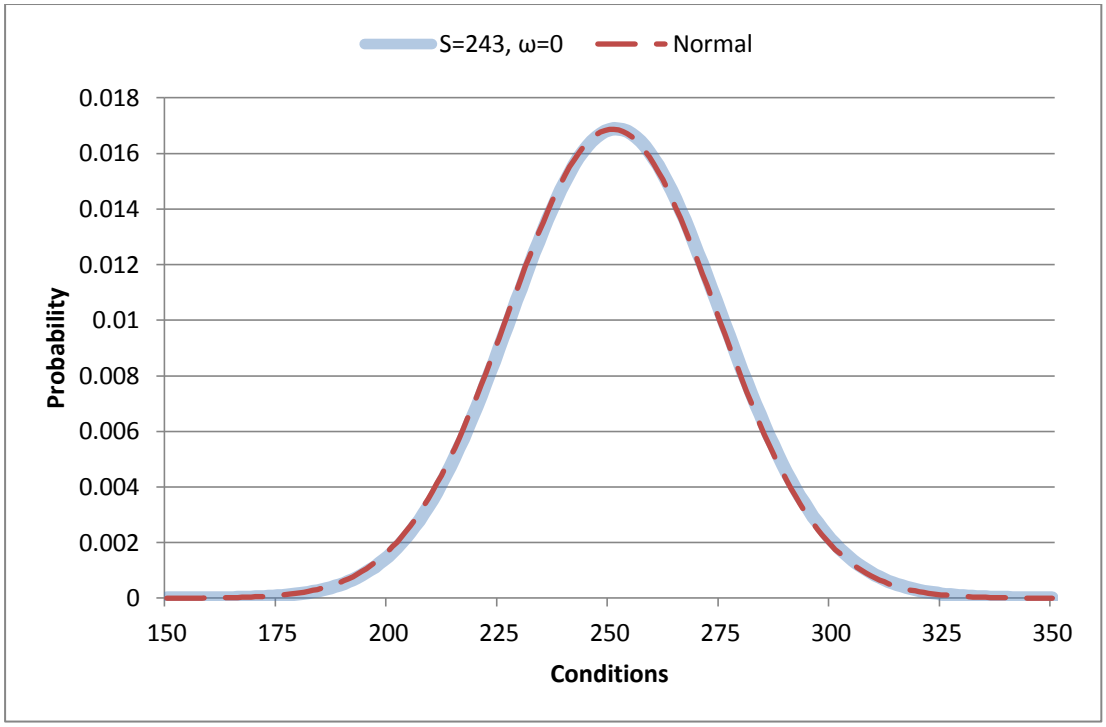


Figure 4.15: Kurtosis of conditions distributions



Despite the systematic skewness and kurtosis shown in Figures 4.14 and 4.15, the numbers involved are very small. We demonstrate in Figure 4.16 that assuming normality causes few problems by plotting one of our conditions distributions along with a normal distribution with the same mean and variance. We choose the situation where Team 1 scores 243 and loses the match, as this is a situation resulting in a relatively high combination of skewness and excess kurtosis and therefore should provide an approximate upper bound of the negative impact of assuming normality. The graph shows that we should not be concerned about assuming normality and the cost of this slight simplifying assumption is likely to be trivial in comparison to the benefits provided by the simulation of a larger number of values for conditions in subsequent analyses. To confirm this, we perform the same normality test that we performed on the first-innings score distribution. That is, we simulate 1000 values of conditions from our posterior distribution and sort the data in ascending order of drawn conditions. If the data are normally distributed the n^{th} score should be equal to the inverse normal of $\frac{1}{n}$, for the simulated mean and variance of conditions. Eliminating the five lowest and five highest observations in order to defend against outliers, the mean absolute deviation of drawn conditions from what would be expected under a normal distribution is 0.7 runs.

Figure 4.16: Implied conditions distribution with normality approximation

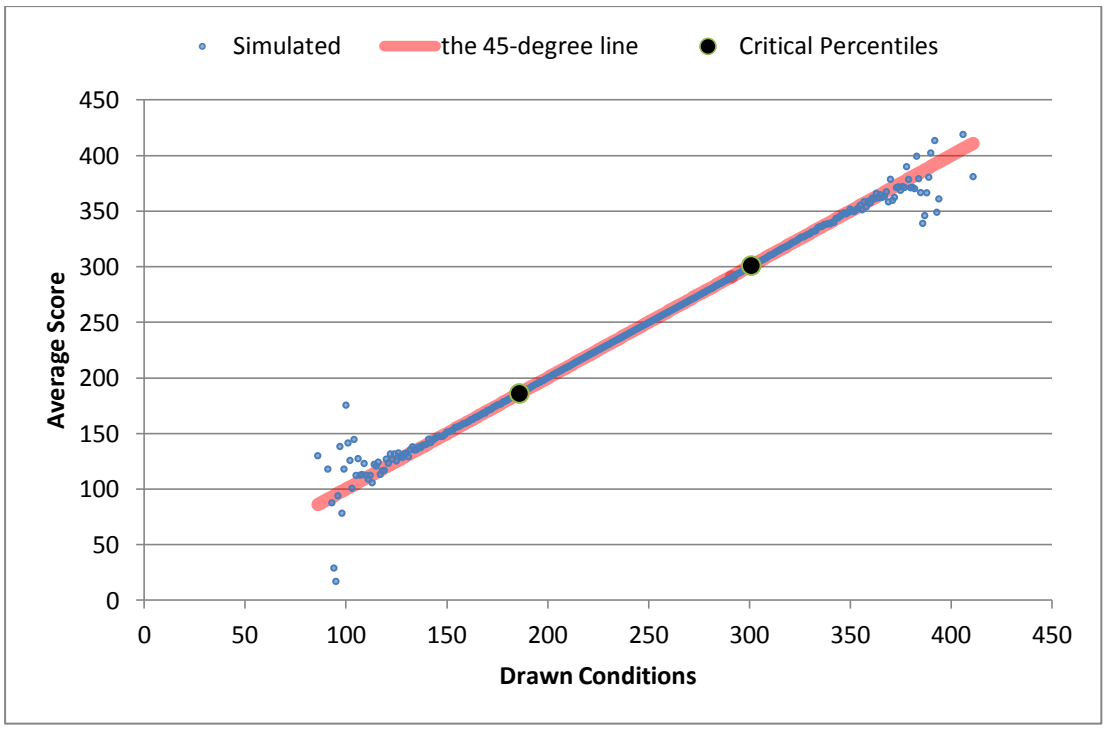


4.4.8 Assessing the fit of the conditional distributions to the data

Theoretically, matches played in conditions with a particular value should result in an average first-innings score of that value. We test our results by employing several Monte Carlo simulations. There are two motivations behind this analysis. It is important to confirm that our method of calculating conditional distributions for conditions and simulating from these distributions for each given score and result actually works. Additionally, it would be useful to know if our data set has any abnormalities that might lead to the average score for each value of conditions not being approximately equal to that value of conditions. This could occur, for example, if an unusual percentage of games had been won by either team around any particular score. This information could help explain any strange results in subsequent analyses using the conditions variable.

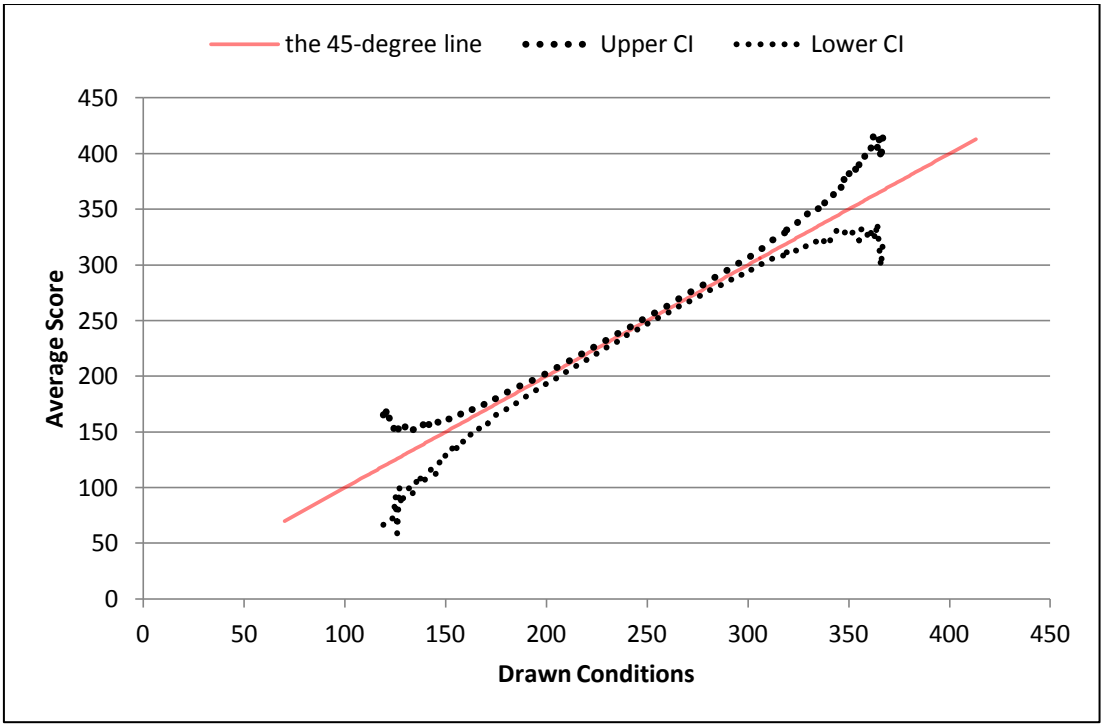
We test the mechanics of our method by randomly drawing one value from the distribution of χ and two values from the distribution of ρ. We add the first draw of ρ to χ in order to determine a first-innings score, *S*, which we round to the nearest integer. If the first draw of ρ is greater than the sum of the second draw plus the performance advantage, this is a win to the team batting first, otherwise it is a loss. We generate 10000 scores and results by repeating these steps. We then can apply the appropriate posterior distribution for conditions to each game and we draw 5000 conditions values from this distribution, again rounding to the nearest integer. This gives us a generated data set with 50000000 observations of score and drawn conditions and we can subsequently determine the average score achieved for each (rounded) value of drawn conditions. We plot the results in Figure 4.17 below, showing the 2.5^{th} and 97.5^{th} percentiles of the overall conditions distributions to show the range of conditions that are most likely to be experienced.

Figure 4.17: Average Score in generated data set



It is clear that the average first-innings score in a given set of conditions closely approximates the value of those conditions. We have, to this point, simply confirmed that our method works in theoretical games and we need to check the relationship between inferred conditions and average first-innings score in our data set of matches. Before doing so, we need to think about the amount of deviation from the 45-degree line that would be acceptable, given our sample size. In order to do this, we randomly sample 784 of the 10000 scores and results previously generated, along with the 5000 draws of conditions for each of those games, and we calculate the average first-innings score for each rounded value of drawn conditions. We repeat this process 100 times, thus generating 100 samples of 784 simulated matches, and generate a 95% confidence interval for the average first-innings score given a particular value of drawn conditions. These confidence intervals are shown in Figure 4.18. Note that we exclude from the confidence interval lines where we did not observe at least one draw of a particular value of conditions in all 100 iterations - that is, where in 784 games and 5000 drawn conditions for each game, we did not observe the particular value of rounded conditions even once.

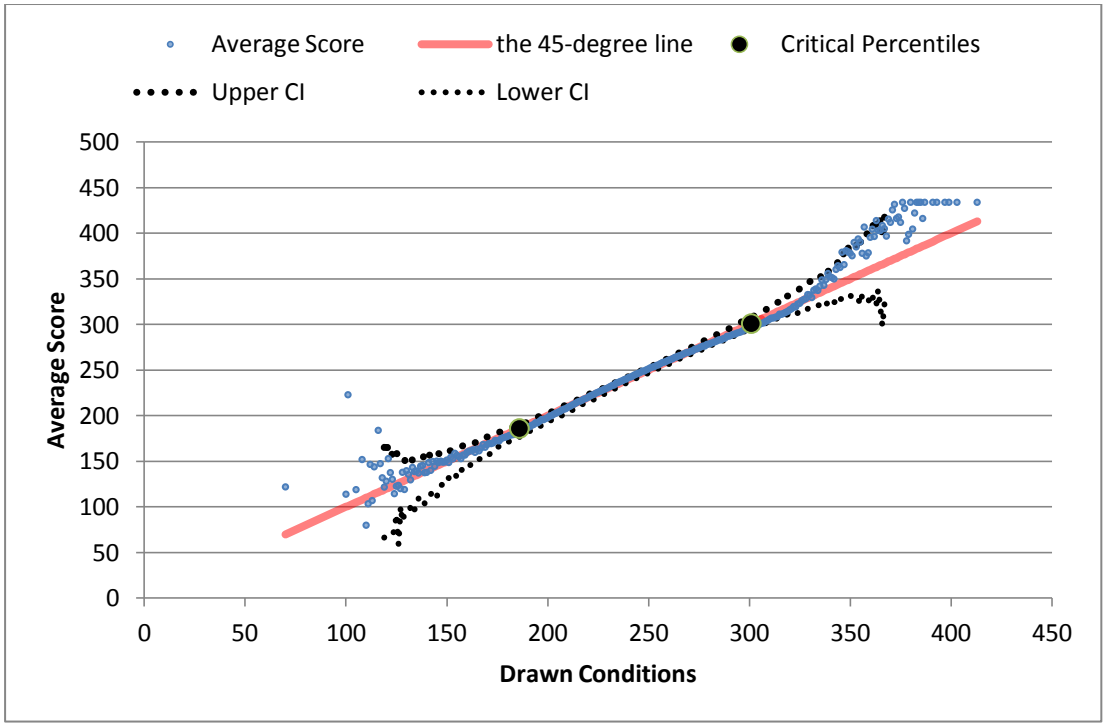
Figure 4.18: Confidence intervals for a sample size of 784



In order to assess the fit of our drawn conditions to the theoretical 45-degree line, we take the actual observed first-innings score and result from our 784 games and apply the mean and variance for the conditions distribution implied by each score and result. As in the previous simulation, we generate 5000 values for conditions from the conditional distribution for each match.

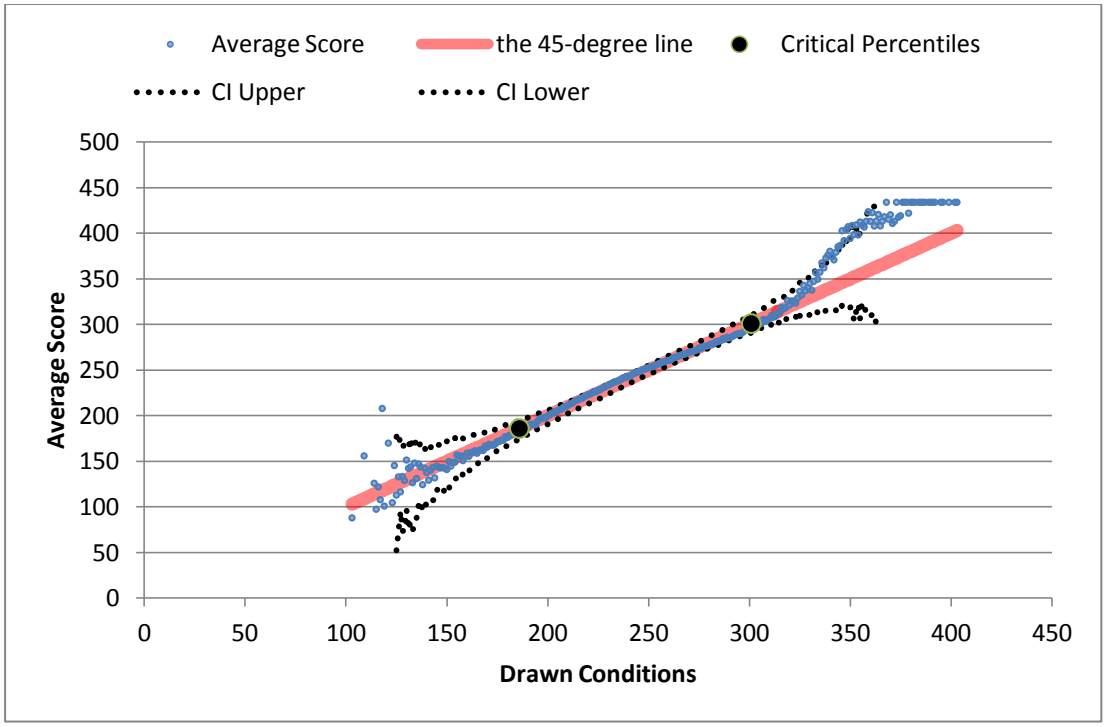
Figure 4.19 shows the average first-innings score for each value of conditions. We see that again the draws from the conditions distributions do a good job of predicting what the average first-innings score will be, particularly within the range in which 95% of conditions fall. The high draws of conditions result in an average score close to the upper bound of the confidence interval over the range that the confidence interval is estimated; this is likely to be due to a particularly unusual game where Australia scored an extremely large total of 434 against South Africa and remarkably lost the match.

Figure 4.19: Average Score in observed data set



For the purposes of determining the various distributions required to calculate the conditional distribution for conditions we used a data set containing games from a ten-year period. This data were easily obtainable as we only required the first-innings score and the result of the game. However, for the analysis in subsequent chapters, we require ball-by-ball data, of which we have a much smaller sample. Our final check is to check the relationship between inferred ground conditions and average first-innings score in these 311 matches for which we have ball-by-ball data. This is shown in Figure 4.20. In this smaller data set we see that there is a large difference between the average score and the inferred conditions at very high values of conditions; however, over the region where 95% of conditions occur, the fit is still very good. Even at the higher draws of conditions, the data are very close to the upper bound of the 95% confidence interval generated from samples of 311 games.

Figure 4.20: Average Score in ball-by-ball data set



4.5 Approach Two: The impact of rule changes

Over the ten-year period of our data set, the rules of One Day Cricket have changed significantly three times. We investigate the data under each set of rules separately. The rule changes predominately concerned the restrictions on where the bowling captain can place his fielders. At the beginning of our data set, the fielding captain could have no more than two fielders outside an oval drawn 30 metres from the wickets for a period of 15 overs at the start of the match. For the remainder of the innings, five fielders were allowed outside the oval. In approximately July 2005,[7] this was reduced to the first ten overs of the match but the bowling captain also had to select two other blocks of five overs in which the restrictions would apply.

[7] At the time of the rule change the old rules were still used for some games for a short period of time.

These blocks of overs are known as “powerplays”. At this time the “supersub” rule was introduced, which would allow each side to make one player substitution at any stage of the game. In March 2006 the supersub rule was cancelled, while the powerplay rule continued. Finally, in October 2008 the powerplay rule was changed to enable the batting side to control when one of the two blocks of powerplay overs was taken.

The increased presence of fielders close to the batsman and the lack of fielders patrolling the boundary serve to increase both scoring rates and the risk of a batsman getting out. There are generally more runs available but it is more difficult to score these runs without hitting the ball over the top of the fielders, rather than along the ground, resulting in the batsman risking hitting a catch. Before we move forward with our analysis, we assume that the minor rule change allowing three fielders in the restricted area during the second powerplay has no significant effect. By far the more significant rule change is the extension of the fielding restrictions from 90 balls to 120 balls in total.

We split the data into four subsets based on the four different sets of rules. It is possible that these rule changes may affect our calculation of the posterior distributions of conditions. Firstly, a change in rules could change the average score, which we set as the mean of the conditions distribution. We note that the average score could change independently of the rules due to, for example, pitches getting easier or harder for batting over time, or teams on average becoming relatively better at batting or bowling. Splitting the data set enables us to include any variation in the average score. It is also possible that a change in the variance of conditions might be observed. This could be due to the ground conditions becoming more variable or the rules allowing for greater or lesser exploitation of the conditions for the batsmen or bowlers

favoured by the conditions. Finally, the rules could impact on the second-innings advantage as they could alter the set of potential risk profiles available to the team batting second.

4.5.1 Data demographics

In Table 4.9 we show the number of matches played under each set of rules. We have far more games under the more traditional 15-over restriction rules, indicating how frequently rule changes have occurred in more recent times. We have the smallest sample size under the combination of bowling-powerplay and supersub rules.

Table 4.9: Number of matches under each set of rules

Rule Number	Description	Number of Matches
1	15-over restrictions	441
2	Bowling powerplay and supersub	58
3	Bowling powerplay only	193
4	Batting powerplay	92

In Table 4.10 we show the matches played by each team under each rule. This enables us to check whether we have introduced any significant team biases in splitting the data set into smaller groups. We see that, other than West Indies batting first under Rule 2, we have a reasonable coverage of each team. Table 4.11 shows that we have a reasonably good distribution of matches played in each country under each set of rules, with the main concerns here being that no matches were played in Sri Lanka or West Indies while Rule 2 was in place.

Table 4.10: Number of matches under each set of rules by team

	Rule 1		Rule 2		Rule 3		Rule 4	
First Innings	**Bat**	**Field**	**Bat**	**Field**	**Bat**	**Field**	**Bat**	**Field**
Australia	66	55	10	7	32	22	22	14
England	43	33	5	3	31	29	11	13
India	55	60	6	13	31	36	14	14
New Zealand	49	60	10	9	16	23	7	11
Pakistan	71	66	6	4	14	23	12	11
South Africa	54	66	4	12	19	21	6	10
Sri Lanka	60	53	16	6	32	15	12	11
West Indies	43	48	1	4	18	24	8	8

Table 4.11: Number of matches under each set of rules by venue country

Country	**Rule 1**	**Rule 2**	**Rule 3**	**Rule 4**
Australia	71	15	27	9
England	43	3	36	9
India	35	11	29	14
New Zealand	45	12	12	4
Pakistan	27	10	13	3
South Africa	64	4	20	22
Sri Lanka	57	0	12	14
West Indies	37	0	29	6
Other	62	3	15	11

4.5.2 Data investigation

We are interested in the accuracy of our assumption of normality of the first-innings scores under the different sets of rules. In Figures 4.21 to 4.24 we plot these distributions. It is clear that when teams scored below the mean under Rule 1 they tended to score further away from the mean than when they scored above the mean. Under Rule 2, the data are fairly unstable, most likely due to the small sample size, while under Rules 3 and 4, the data look reasonably close to normal. Overall, we reiterate our earlier point that normality seems to be the best choice out of the set of functional forms from which we could reasonably choose, and

we again note that any errors arising from imperfect normality are likely to be small compared to the error of ignoring conditions altogether in our analyses in subsequent chapters.

Figure 4.21: First-innings scores under Rule 1

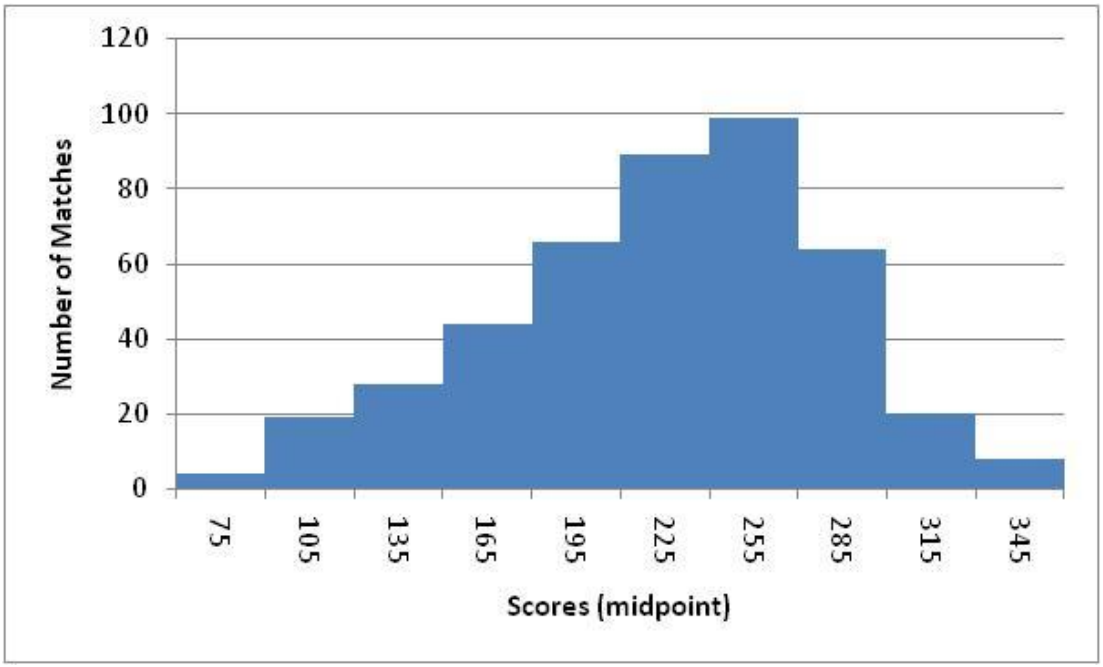


Figure 4.22: First-innings scores under Rule 2

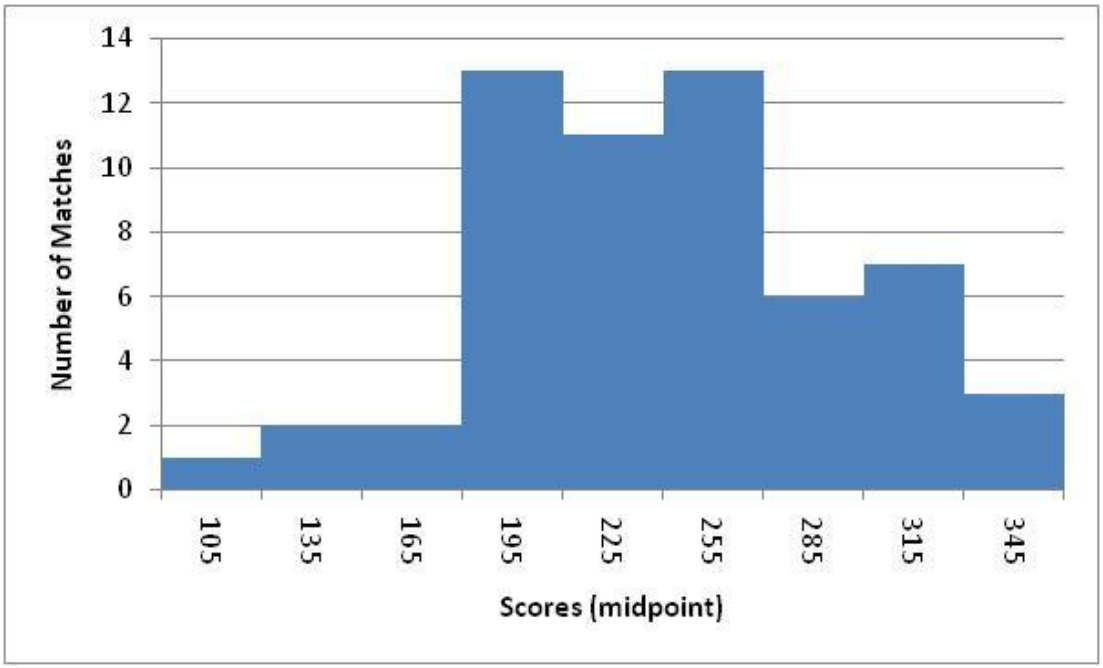


Figure 4.23: First-innings scores under Rule 3

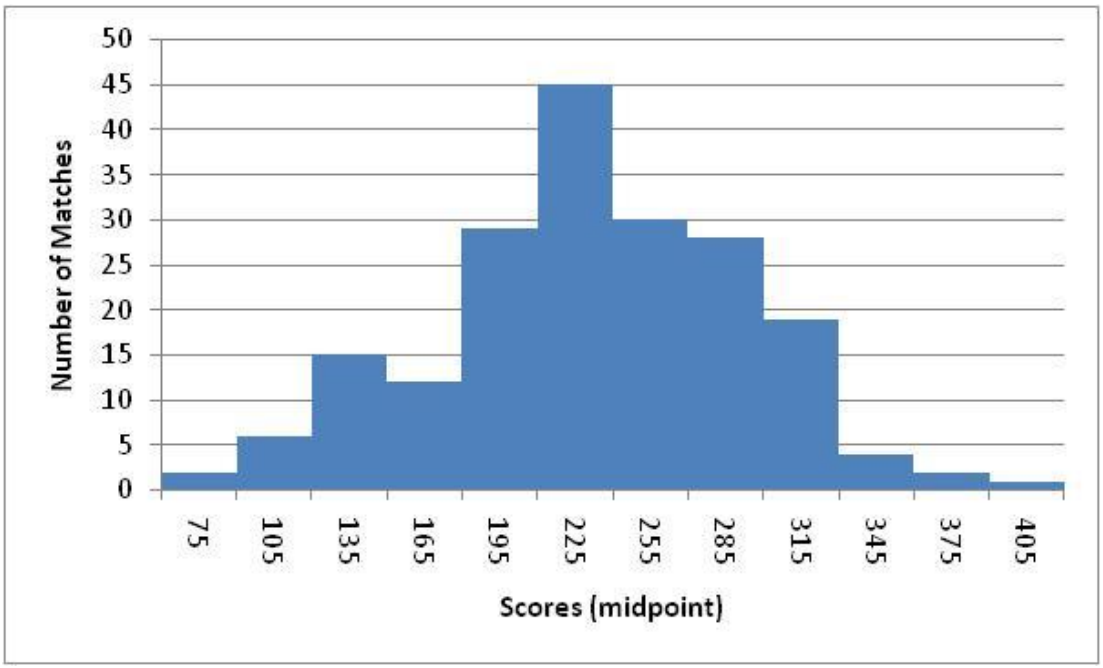
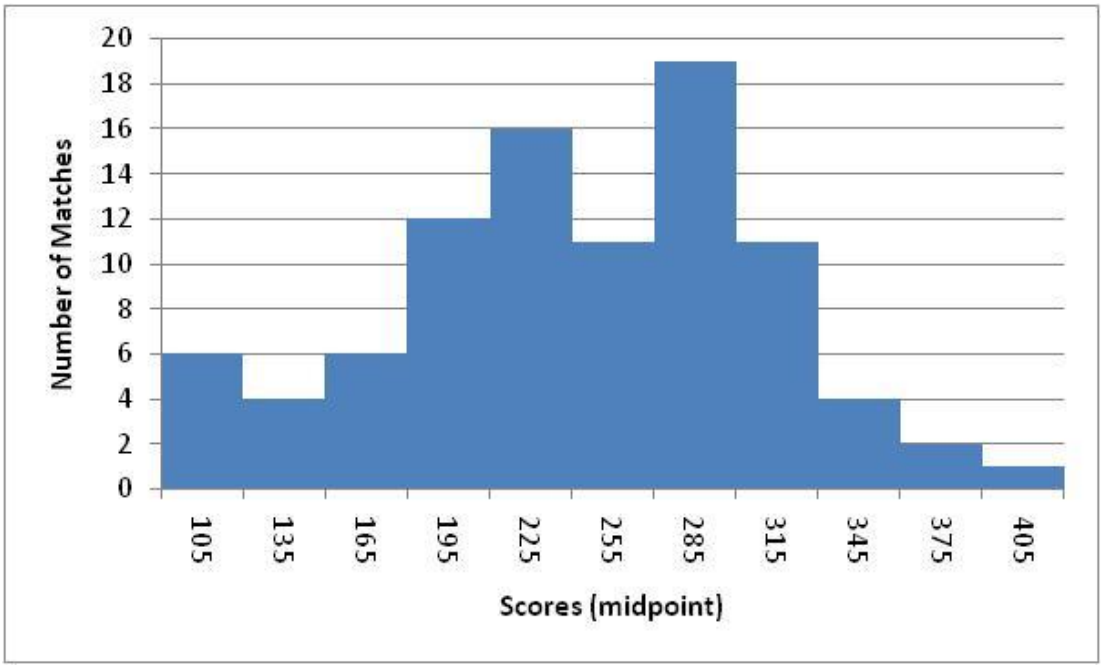


Figure 4.24: First-innings scores under Rule 4



There are some outcomes that we expect under the various rules, based on some cricket intuition. First, as the power-play rules are restrictions on the bowling team, there should be no

disadvantage to batting teams and therefore the first and second-innings means should be higher under Rules 2, 3 and 4 than under Rule 1. Second, we hypothesise that batting sides will be able to take advantage of the restrictions by scoring more quickly when they are in good positions and there should be no effect when the batting side is in a bad position (since nothing prevented the fielding captain from bringing up the field in the absence of the power-play rule). This means that the first-innings variance should be higher under Rules 2, 3 and 4 than under Rule 1. Third, there is no reason to believe that the variance of conditions should have changed as a result of the introduction of each rule, although it may change exogenously over time. Finally, the second-innings variance should be at least as great as the first-innings variance. At a minimum, Team 2 should be able to completely ignore their target score and play as if they were simply a score-maximising team in the first innings. Even with a "dumb" Team 2 and zero conditions variance model, the variances of the first and second-innings scores should be approximately equal.

In Table 4.12 we present the mean and variance of the first-innings score for each of our four separate datasets, as well as the second-innings mean and variance implied by a Probit model of the result versus the first-innings score.

Table 4.12: Mean and variance of first-innings score under different rules

Statistic	Rule 1	Rule 2	Rule 3	Rule 4
N	441	58	193	92
μ_S	238.104	255.741	245.596	255.435
σ_S^2	3114.366	2730.230	3712.242	4371.831
μ_{S_2}	240.680	276.501	253.261	257.988
$\sigma_{S_2}^2$	4736.622	6805.365	9553.529	2391.888

This table presents some very interesting results. As expected, Rules 2, 3 and 4 seem to have significantly favoured batting, with Rule 2, the powerplay and supersub combined rule, having been particularly easy to bat second under. However, we note that the variance behaviour is not as expected, with the second-innings variance changing significantly from rule to rule. While we accept that it is perfectly reasonable that the second-innings variance should change under different rules, we note that under Rule 4, the second-innings variance is substantially lower than the first-innings variance.

We do not believe that there are cricket reasons to explain the difference observed in Table 4.12; therefore, we conclude that the most likely scenario is that the sample sizes are too small for the observed data to closely approximate reality. We note that we have a further option of combining Rules 2, 3 and 4 and analysing the data as two data sets. The logic behind this approach is that, while rules 2, 3 and 4 are all different, they have one common factor that is lacking in Rule 1 - that is, the use of 20 overs of fielding restrictions in total, rather than 15 overs under Rule 1. We need to decide whether we will use this alternative split, or simply leave the data as one set.

We define the matches played under Rule 1 as “Non-Powerplay” matches and all other matches as “Powerplay” matches. Table 4.13 shows the summary statistics for each of these data sets as well as a reminder of the summary statistics of the full data set. We also show the difference for each parameter between the powerplay and non- powerplay data sets.

Table 4.13: Mean and variance of first-innings score under different rules

Statistic	Full Data Set	Non-powerplay	Powerplay	Difference
N	784	441	343	N/A
μ_S	243.287	238.104	249.950	11.846
σ_S^2	3412.488	3114.366	3726.848	612.482
μ_{S_2}	247.981	240.679	257.911	17.232
$\sigma_{S_2}^2$	5673.117	4736.585	6418.007	1681.422
δ	75.119	79.337	73.473	-5.864
A_ρ	3.526	2.042	5.849	3.807
σ_ρ^2	2563.412	2470.854	2738.125	267.271
σ_χ^2	849.076	643.512	988.633	345.121

We see from Table 4.13 that the difference between the powerplay and non-powerplay data is fairly substantial for all parameters. The large increase in the variance of conditions is not explained by our earlier intuition, although as mentioned conditions may have changed exogenously. In order to determine the best way to proceed with the two data sets, we undertake a Monte Carlo study to assess the significance of each difference. Our null hypothesis is that all the data are from the same population. We use the parameters estimated in the full data set and proceed as follows.

1. Draw 441 values of ρ_1 and χ for the first-innings performance and conditions, respectively.
2. $S = \rho + \chi$ gives 441 first-innings scores.
3. Draw ρ_2 for each game to represent second-innings performance.
4. If $\rho_1 > \rho_2 + A_\rho$ then $\omega = 1$, else $\omega = 0$.
5. Run a Probit model of ω on *S*.
6. Calculate μ_S, μ_{S_2}, σ_S^2, $\sigma_{S_2}^2$ and δ.
7. Repeat steps 1. to 6. but drawing 343 values this time.
8. Calculate the difference between each of the parameters in step 6. in the two data sets.
9. Repeat steps 1. to 8. 10000 times.

We now have a distribution for the difference in each parameter under the null hypothesis that the two samples are from the same population. Looking at one parameter at a time, we sort our simulated distribution by the parameter in question and look up our observed difference from Table 4.13 in this simulated distribution in order to see at which percentile of the simulated distribution the parameter occurs. If the parameter is more extreme than the 50^{th} or 950^{th} observation in the simulated distribution, then we have some evidence at the 10% significance level to reject our null hypothesis that we have one big population. The results are shown in Table 4.14.

Table 4.14: Position of observed difference in simulated distribution

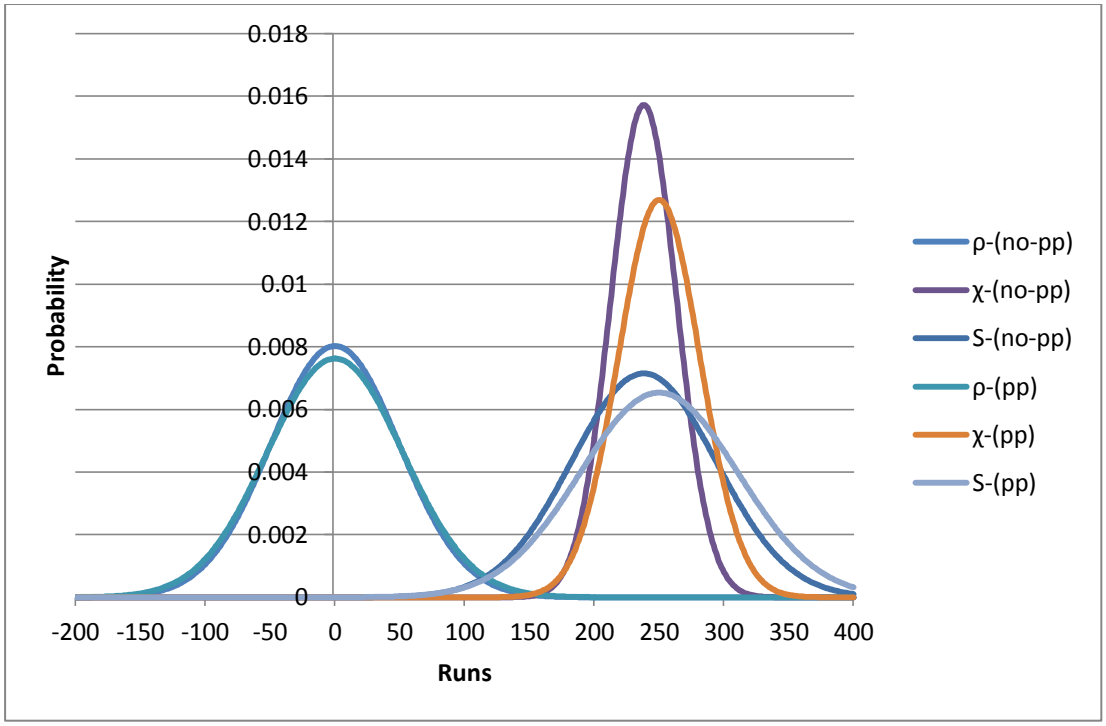
Statistic	Difference	Position in simulated distribution	P-Value
μ_S	11.846	9974.5/10000	0.0051
σ_S^2	612.482	9613.5/10000	0.0773
μ_{S_2}	17.232	9891.5/10000	0.0217
$\sigma_{S_2}^2$	1681.422	8217.5/10000	0.3565
δ	-5.864	3539.5/10000	0.7079
A_ρ	3.807	7219.5/10000	0.5561
σ_ρ^2	267.271	8565.9/10000	0.2869
σ_χ^2	345.121	7532.5/10000	0.4935

Based on the results shown in Table 4.14, we can be very confident that the powerplay data set has a higher mean in both innings and confident that it has a higher first-innings variance. While the difference in the second-innings variance is rather large, we do not find significant evidence that this difference would not have been observed from two data sets drawn from a single population. There is simply a substantial amount of uncertainty in

determining the second-innings variance. The parameter δ represents the percentage of first-innings variance that is attributed to performance. We do not find sufficient evidence of a change in this relationship between the data sets and subsequently we do not find substantial evidence of a change in the variance of performance or conditions. Despite the uncertainty around the second-innings variance, it nevertheless represents the best possible available estimate and we therefore decide to proceed with our splitting of the data set on the basis that the means of the data sets in both innings are significantly different.

We plot the assumed first-innings performance, conditions and score distributions in Figure 4.25. We see that, while there has been very little change in the shape of the performance distribution, conditions and scores have become more variable in the powerplay data set. Note that while the overall average score has risen, we cannot know whether it is because conditions have become easier to bat in or if average batting performance has improved, relative to average bowling performance. We display the data as if it is the former; however, this is by construction rather than by analysis since we have defined performance to be the deviation from the average score expected in the conditions.

Figure 4.25: Performance, Conditions and Score prior distributions



4.5.3 Selected results

We show some examples of the posterior distributions of conditions given the first-innings score and the result of the match. The focus here is on showing the difference between the posterior distributions calculated for the non-powerplay dataset, those calculated for the powerplay dataset and those calculated for combined dataset. Figure 4.26 shows a match with a score equal to the overall dataset mean of 243, with a win to Team 1. Figure 4.27 shows a low score of 200 with a win to Team 2, and Figure 4.28 shows a high score of 300 with the unexpected result of a win to Team 2. Tables 4.15, 4.16 and 4.17 show the summary statistics for each situation.

Figure 4.26: Inferred conditions with a mid-range score

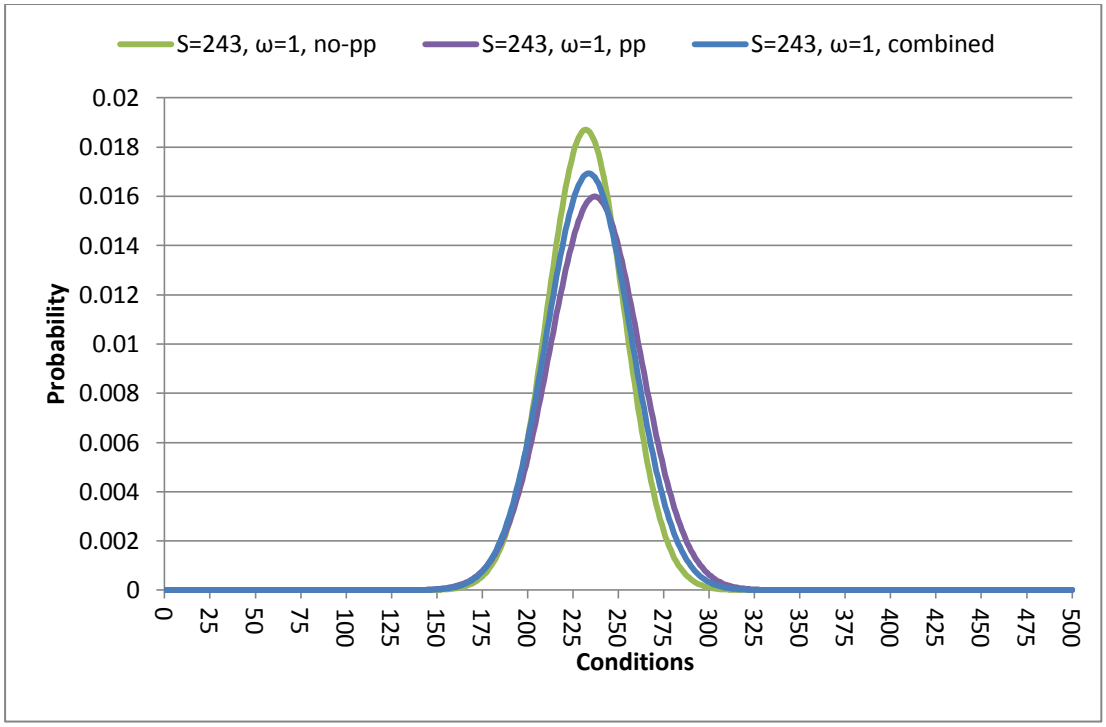


Table 4.15: Mean and variance of conditions with mid-range score

Conditions Distribution	Mean	Variance
$S = 243, \omega = 1, no - pp$	231.855	455.536
$S = 243, \omega = 1, pp$	236.753	623.547
$S = 243, \omega = 1, combined$	233.741	555.113

When we look at mid-range scores, we see a higher mean and higher variance in the conditions inferred by the powerplay data set and we are more confident about the value of conditions in the non-powerplay data set.

Figure 4.27: Inferred conditions with extreme score, expected results

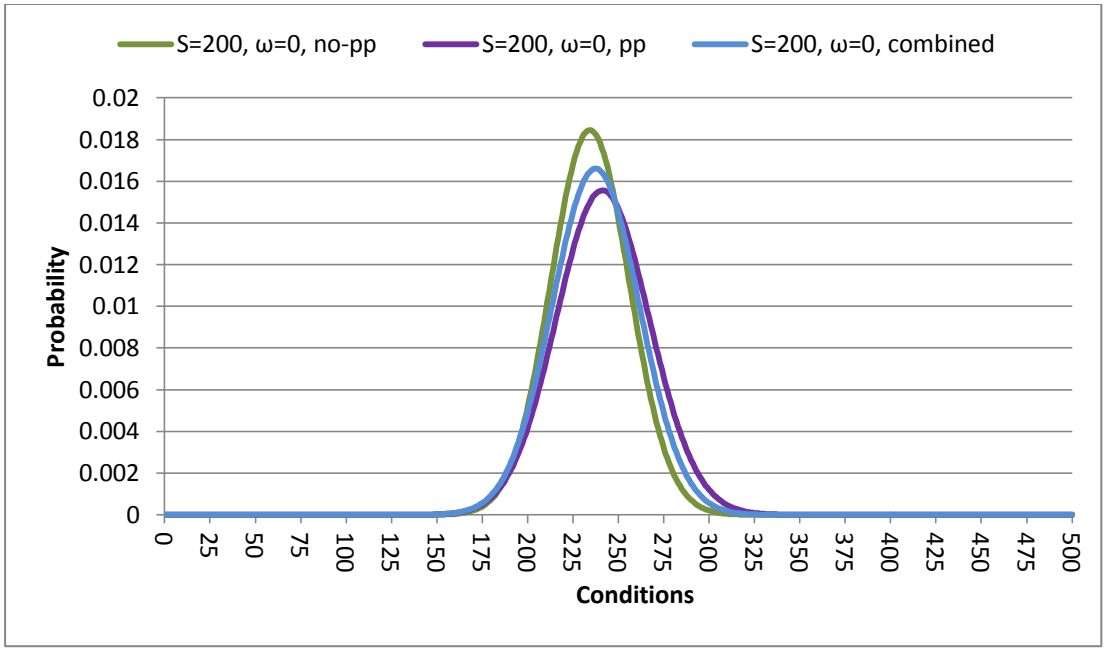


Table 4.16: Mean and variance of conditions with extreme scores, expected results

Conditions Distribution	Mean	Variance
$S = 200, \omega = 0, no - pp$	234.565	467.811
$S = 200, \omega = 0, pp$	241.655	657.621
$S = 200, \omega = 0, combined$	237.491	577.335

Despite extreme first-innings scores, the results going the way that they would be expected to go indicates that the means of the posterior distributions are not far away from the overall means.

Figure 4.28: Inferred conditions with extreme scores, unexpected results

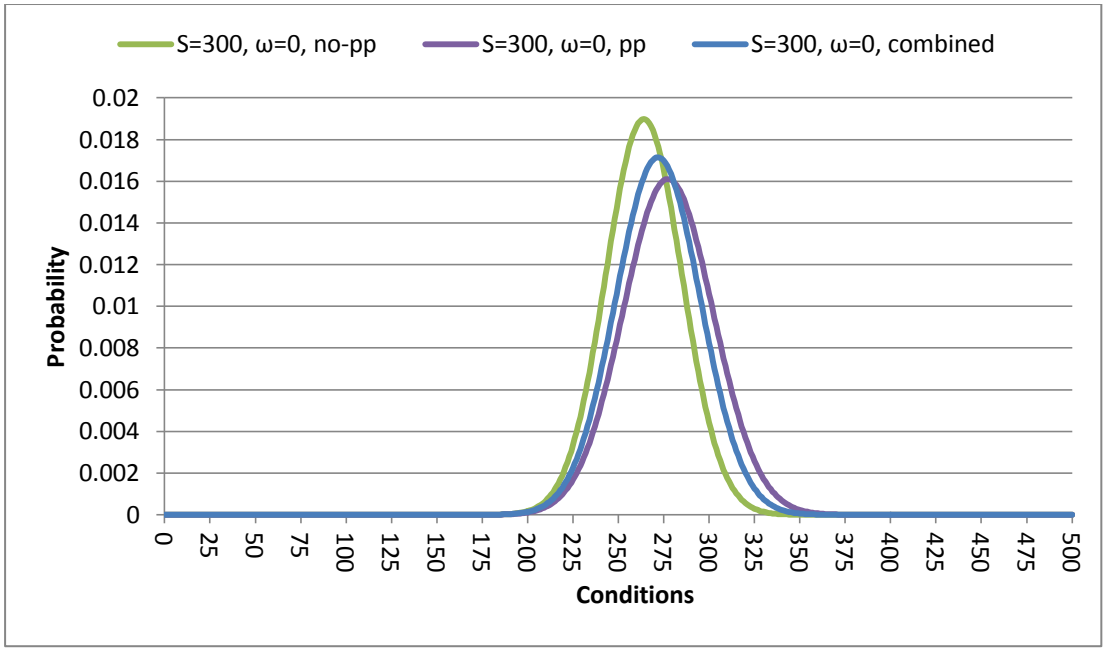


Table 4.17: Mean and variance of conditions with extreme scores, unexpected results

Conditions Distribution	Mean	Variance
$S = 300, \omega = 0, no - pp$	264.131	442.004
$S = 300, \omega = 0, pp$	277.538	614.349
$S = 300, \omega = 0, combined$	271.865	541.146

When extreme scores are scored and the results that occur are unexpected, then we have far greater evidence that conditions are not close to the overall mean. We note that even with a very high first-innings score and an unexpected result, the mean of the posterior distributions for conditions are still some way short of the observed first-innings scores. This fits with the idea that performance is the dominant factor in determining first-innings score.

4.5.4 Assessing the fit of the conditional distributions to the data

As we did when we analysed the data as one complete data set, we simulate from the inferred conditions distributions for each game and plot the observed average score for each rounded value of conditions. In Figure 4.29 we show the full data set of 784 games and in Figure 4.30 we show the 311 games for which we have ball-by-ball data.

Figure 4.29: Average Score in split analysis of observed data set

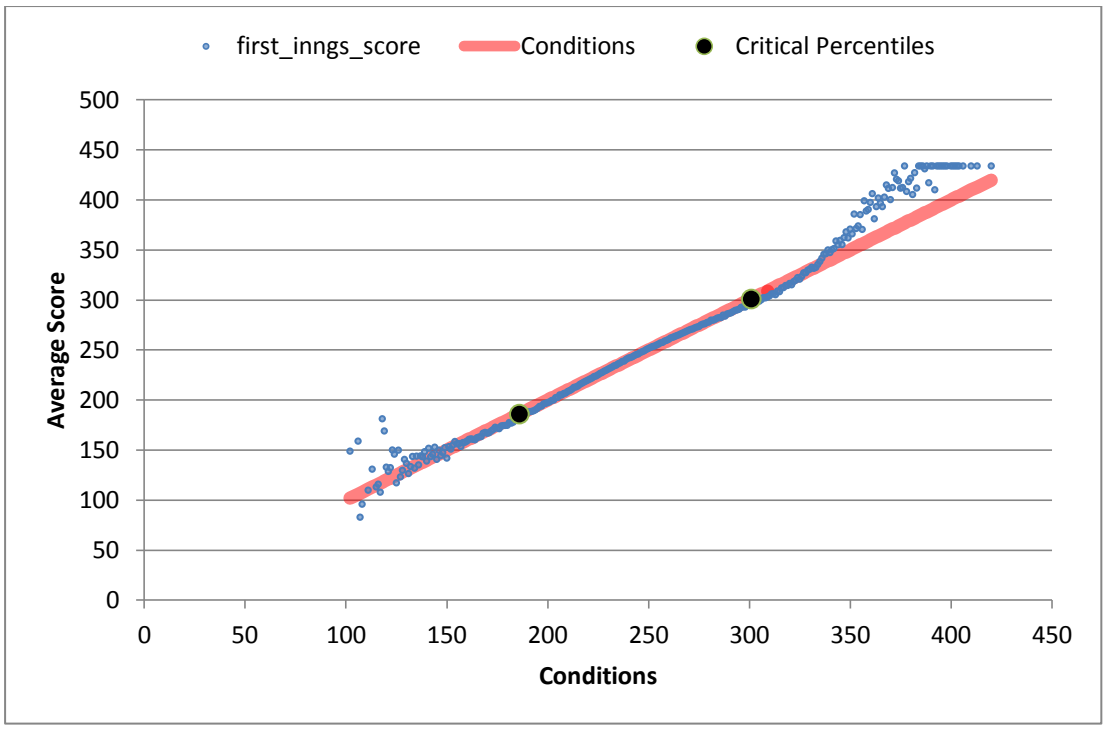
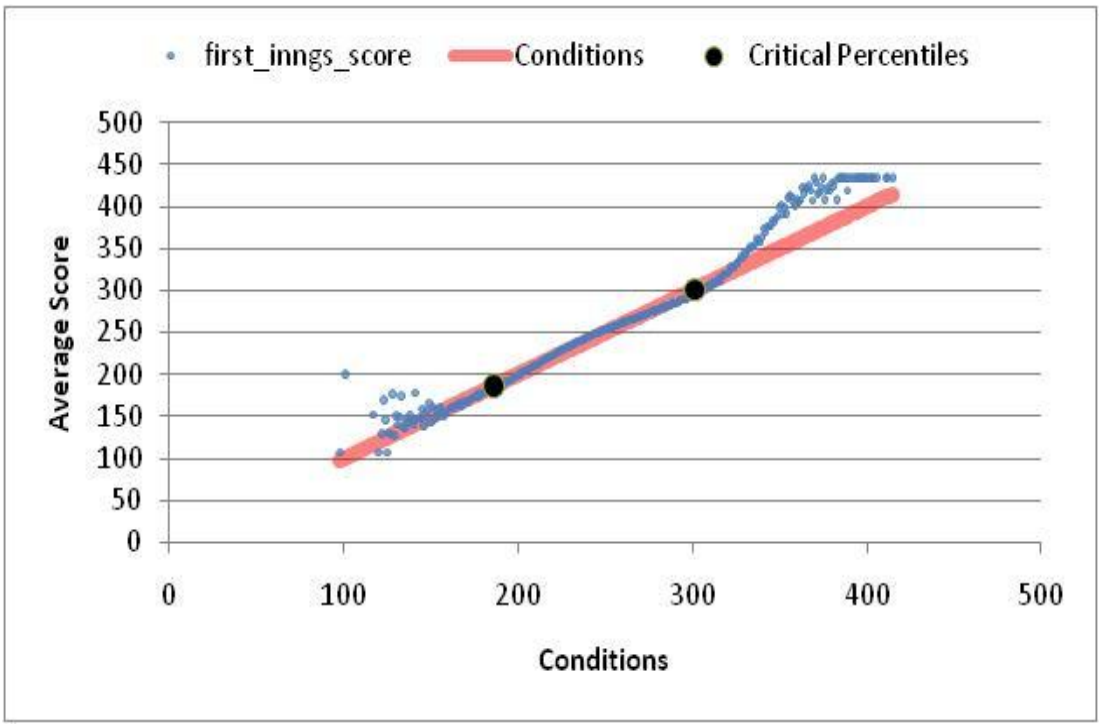


Figure 4.30: Average Score in split analysis of ball-by-ball data set



The data from the split analysis seems to be a similar fit to when we simply used the one data set.

4.6 Conclusion

By assuming a functional form for a model of first-innings score, determining the contribution to the total score variance of each component in the model and applying Bayes' Theorem, we have obtained information pertaining to a critical but unobservable variable. This information is in the form of a distribution that is conditional on the first-innings score and the result of the game. In our subsequent chapters, we are able to randomly draw values from these conditional distributions in order to include conditions as a right-hand-side variable, greatly enhancing the predictive ability of our other models.

CHAPTER 5

A first-innings dynamic programme

5.1 Introduction

In the first innings of a game of ODI cricket, cricket fans commonly ask questions of each other such as "what total will they end up with from this position?" or "do you think they will get enough to win?" In this chapter, we build a model of the expected future outcomes from the first innings. Such a model is useful not only for prediction, the focus of this chapter, but also for estimating the trade-off that each batsman faces between scoring rate and survival rate and testing the strategic optimality of a batsman. This involves the creation of Production Possibility Frontiers (PPFs) for individual batsmen and this is the focus of Chapter 6.

Cricket is a difficult sport in which to speculate about future outcomes while a match is in progress. This is, for the most part, caused by the sequential nature of the game. Team 1 bats for an innings, sets a target, which Team 2 then attempts to beat. This is in stark contrast to simultaneous games such as any of the football codes, where the two teams are each attempting to score and therefore the current score is a good indicator of which team is more likely to win the match.[8] The fact that cricket is a sequential game means that we need something other than the current score of the other team with which to assess Team 1's progress.

[8] We say a good indicator, but not a perfect indicator as there may be other information available, such as; the weaker team is currently winning, or the team currently winning has had the better of the conditions, which will change at half time when the teams swap ends of the field.

A dynamic programming model of the first innings is one way to predict the likely total score of Team 1, from any given position. This predicted total is a very good indicator of how well Team 1 has played so far in the innings, as it indicates the size of the task that Team 2 will be likely to face later in the match. We construct our models both with and without the conditional distribution of the ground conditions variable, χ , calculated in Chapter 4, in order to assess the impact of this variable on our models.

Note that we consider only the first innings in this chapter as we are able to construct a model of expected future outcomes without considering the current or target score, which would greatly expand the state space. Using a model with fewer state variables means that we have more data in each state, making for more reliable PPF estimation in Chapter 6. A second-innings dynamic programme is constructed in Chapter 7, for the purpose of a specific second-innings application, namely, predicting the winner of rain-interrupted matches.

Our dynamic programming model will enable us to define an important new variable, the cost of a wicket. This represents the difference in Team 1's expected total score in the current situation and in the situation where they have lost an additional wicket. This new variable is an important input to Chapter 6 as it enables us to infer the likely strategy of individual batsmen and subsequently construct production possibility frontiers (PPFs) of the trade-off between scoring rate and survival rate for selected players. We show that these PPFs can be used to compare batsmen, in terms of their ability, their suitability to different match situations and their strategic nous.

5.2 The structure of the model

The objective function of the dynamic programme is simple. The goal of any team in a game of ODI cricket is to maximise the probability that it wins the game.[9] Each team is therefore trying to maximise its probability, π, of winning the game. We write the objective function as

$$\pi = \Pr(Win)$$

and our optimisation problem as

$$Max(\pi)$$

There are a very large number of factors that could affect a game of cricket; however, we need to restrict the scope of our analysis in the interests of parsimonious model building. The main factors describing the state of a game in the first innings are as follows:

1. The number of balls that have been bowled in the innings;
2. The number of wickets that have been lost in the innings;
3. The ground conditions;
4. The number of runs scored in the innings;
5. The ability of the two batsmen current batting;
6. The ability of the batsmen sill waiting to bat (if any);
7. The ability of the bowler currently bowling (if an over is in progress);
8. The ability of the bowlers available to bowl and the number of overs each has remaining;

[9] There may be rare exceptions, in a multi-stage tournament or league where a team needs to win a game by a particular margin in order to get ahead of another team and qualify for the next round. Alternatively, it might be the case that the team in question simply has to avoid a heavy loss to qualify. We believe that these rare exceptions would not create significant bias.

The last four points involve the skill sets of individual players. To take these factors into account would require the creation of millions of different models, one for every possible combination of these four factors. This is very impractical; however, we are interested in investigating average performance. We build these models on the basis of the outcomes that we would expect from two average teams playing against each other. We expect that teams only select bowlers who are at least reasonable bowlers and we assume that the number of wickets lost is a good proxy measure of the skill of the current batsmen and the batsmen still to come. In certain models, we also exclude the impact of the ground conditions, in which case we are looking at the outcomes that we would expect from two average teams playing against each other in unknown ground conditions. In order to create a more useful model, we only include matches played between two top-eight-ranked teams in world cricket, in order for an assumption that the average model approximates reality for any team to be reasonable.

We assume that, in the range in which first-innings totals generally occur, there is a linear relationship between the first-innings score and the probability of winning. This means that an extra run is equally valuable regardless of the final score. For example, a score of 261 gives the team batting first the same advantage over a score of 260 as the advantage that a score of 231 would give them over a score of 230. We show evidence of this later in this chapter. It is also relevant that, even in the presence of non-linearity in the relationship between first-innings score and the probability of winning, on any given ball of the 300 balls in the innings the performance of a team can only influence their expected score by a small amount. A local linear approximation to the true relationship between first-innings score and the probability of winning is unlikely to lead to substantially different decisions. The implication here is that a

team should maximise their expected additional runs, for the vast majority of possible situations that they could be in. We are effectively making current score irrelevant to future decision making. This enables us to revise our objective function for the first innings.

Let V be the expected additional runs to be scored from a particular first-innings situation. The first-innings optimisation problem under the assumption of this linear relationship between score and the probability of winning the match is

$$Max(V)$$

Finally, we make the assumption that future performance within a game is independent of the performance in that game so far; that is, we are assuming no serial correlation of the outcomes on previous balls to the outcome on the current ball. This is a reasonable assumption since the attitude towards risk of each team should be focused around the number of balls and wickets remaining. We note that there may be some serial correlation due to small differences in team ability or a bowler or batsmen playing particularly well on a particular day; however, we believe that our average models will still provide a very good indication of the likely future outcomes of the match.

These simplifying assumptions enable us to define the remaining state variables.

Let i be the number of the next ball in the innings,
where the 301st ball indicates that the innings is complete $i \in \{1, 2, ..., 300\}$

Let j be the current number of wickets lost by the batting team, $j \in \{0, 1, ..., 10\}$

Let χ be the value of the ground conditions, $\chi \in (0, \infty)$

We are now ready to proceed with our estimation of the first-innings dynamic programme.

5.3 The data set

In addition to the first-innings score and result data that we used for the majority of Chapter 4, we have ball-by-ball information for a subset of 311 matches played over the period 20 July, 2001 to 25 January, 2008. These data were collected by New Zealand Cricket and they simply collected the data from as many games as possible, rather than by using any particular sampling strategy. All the matches in the data set are between two top-eight-ranked countries - that is, Australia, England, India, New Zealand, Pakistan, Sri Lanka, South Africa and West Indies.

Table 5.1 shows the number of matches in which each team has batted first, batted second and been the host country. This is to check that we have a reasonable variety of the different situations. Team 1 refers to the team batting first, while Team 2 refers to the team batting second. There are a relatively large number of games where New Zealand is Team 2, but these games still only make up 21% of the overall data set so we do not believe that this will cause any significant bias in our average models.

Table 5.1: Distribution of team and venue information

Country	Team 1	Team 2	Host
Australia	50	30	52
England	33	39	42
India	40	43	32
New Zealand	47	65	40
Pakistan	43	35	28
Sri Lanka	43	32	41
South Africa	34	35	35
West Indies	21	32	27
Other	-	-	14

There was a significant rule change over the period of our data set as the power-play rules were introduced in July 2005. As this rule does affect the way that we create the dynamic programmes, we note that we have 185 "non-power-play matches" and 126 "power-play matches". None of our matches occur during the period since the batting power-play rule was introduced.

5.4 Testing the linearity assumption

We have now made all the assumptions required to estimate a reasonably parsimonious dynamic programme. Before we do so, we need to test our earlier assumption that the probability of winning is linear in the first-innings score, *S*. This assumption enabled us to define the objective function for the team batting first as the expected additional runs from any point. In order to test this assumption we look at the relationship between actual first-innings scores and the percentage of games won with each score. Since we might have very few observations (in some cases no observations) at each score *S*, we need to smooth the data. Our

method is to look at a range of scores in the vicinity of *S*. We use the 41-point interval {*S-20, … ,S+20*} and calculate the percentage of games won by Team 1 in this interval and we repeat this analysis for each value of *S*.

Since only scores and results are required for this analysis, we use the full data set from the decade of the 2000s as described in Chapter 4. We split our data set into two parts: those games played prior to the power-play rule change (non-power-play) and those games played after the rule change (power-play). Figure 5.1 shows the relationship between first-innings score and the percentage of games won in the non-power-play era, while Figure 5.2 shows the relationship for the power-play era. We include a 95% Wilson confidence interval for the estimated proportion of wins, as recommended by Brown et al (2001).

Figure 5.1: Smoothed win percentage versus score in non-power-play data

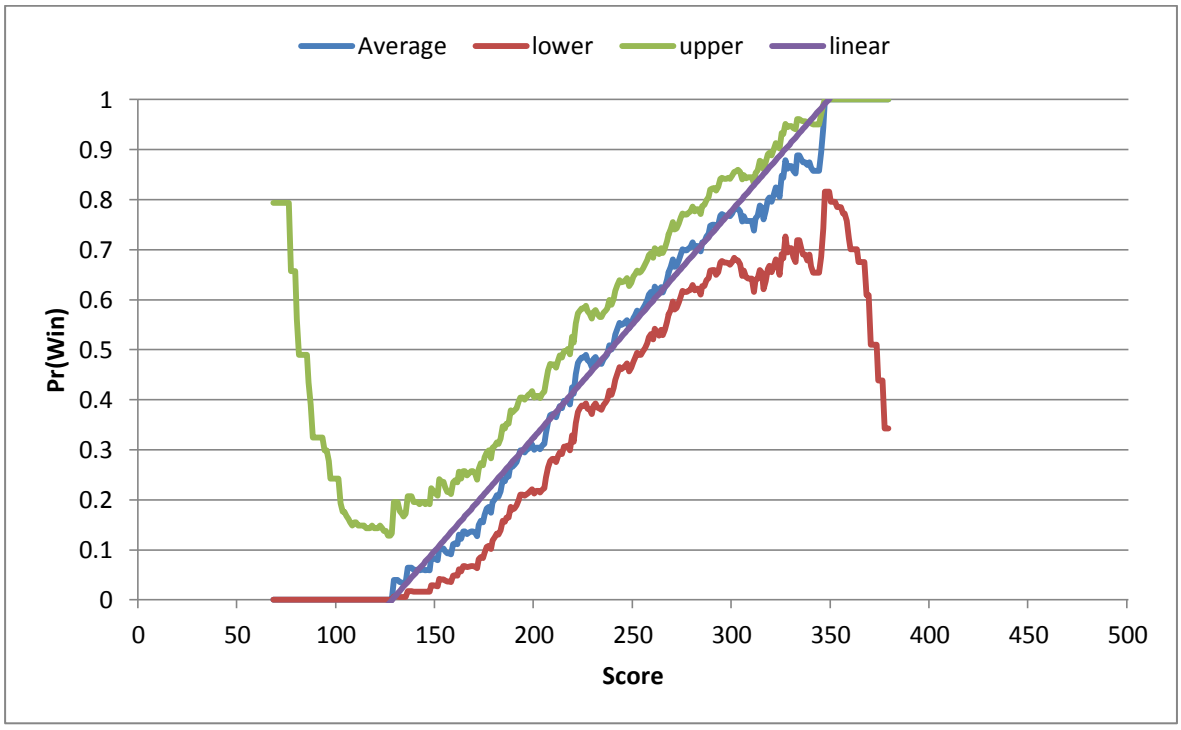
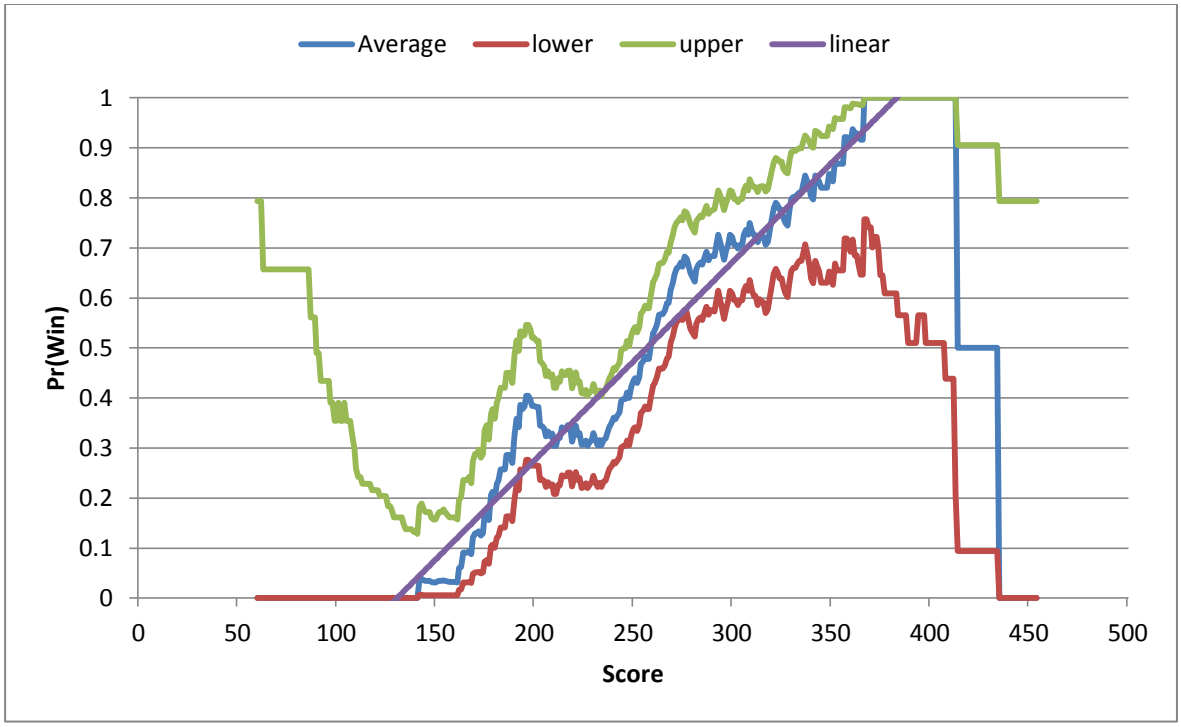


Figure 5.2: Smoothed win percentage versus score in power-play data



It is apparent that assuming a linear relationship between first-innings score and the probability of winning is appropriate for the non-power-play data. Any deviations of the linear model from the observed win percentage are well inside the confidence intervals for the majority of observed scores. In the power-play data, the linearity assumption does not fit the data as well. Two aspects of Figure 5.2 are of particular note. There is an unexpected decreasing trend in the range of scores (197, 234) and the win percentage falls away to zero, rather than the intuitive level of one, when scores get extremely high. The latter situation is because the sample size is one at these points; the game with the highest score simply happened to result in a loss for the team batting first. The decreasing trend is more difficult to explain but a possible cause would be if conditions were worth relatively low amounts over this range when compared to the scores. When we incorporate conditions into our models we implicitly

make this assumption as our conditional distributions for conditions have a lower mean when the game was lost by the team batting first, given a certain value of first-innings score.

5.5 The first-innings value function

We develop a value function, *V*, to calculate the expected additional runs for every possible state.

Let $V(i, j, \chi)$ be the expected number of additional runs given *i* and *j*, $V(i, j, \chi) \in [0, \infty)$

Note that the value of the ground conditions, χ is technically a state variable as it describes the prevailing conditions on the day of the match; however, since we assume that it remains constant throughout the match, it enters the model as a parameter in the functions that determine run scoring and wicket loss. It is therefore not necessarily to consider χ a state variable in the dynamic programme. Additionally define the following:

Let χ be the value of the ground conditions, $\chi \in [0, \infty)$ [10]

Let r_{ij} be a random variable indicating the number of runs scored on legitimate ball *i* given *j*, $r_{ij} \in \{0, 1, .., \infty\}$

Let λ_{ij} be the probability of losing a wicket on ball *i* given *j*, $\lambda_{ij} \in [0, 1]$

Let γ_{ij} be the probability of a wide or no-ball being bowled on legitimate ball *i* given *j*, $\gamma_{ij} \in [0, 1]$

[10] Technically, as χ is drawn from a normal distribution, it is not bounded at zero. However, we are defining it as being censored at zero to fit with the cricketing reality of scores necessarily being non-negative. The mean and variance of the normal distribution make this distinction empirically irrelevant.

Let τ_{ij} be the total number of runs scored from a single occurrence of a wide or a no-ball given *i* and *j*, $\tau_{ij} \in \{1, 2, .., 7\}$

The Bellman equation is defined as follows:

$$V(i, j) = E[r_{ij}] + \lambda_{ij} V(i+1, j+1) + (1-\lambda_{ij}) V(i+1, j) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}} \quad (19)$$

The terminal points, representing the end of the innings, are defined as

$V(301, j) = 0$, for all *j*

$V(i, 10) = 0$, for all *i*

In words, Equation (19) is saying that the expected additional runs scored by the batting team from their current state of being at the i^{th} ball of their allotted 300 and having lost *j* wickets of their allotted ten is equal to expected runs from the next ball *plus* the value function applicable on the next ball *plus* the expected runs scored from extras. The next state is always one of two possible states: one more ball and one more wicket than the current state (with probability λ_{ij}) or one more ball and the same number of wickets as the current state (with probability $(1-\lambda_{ij})$). Note that the final term is the infinite sum of a geometric series as a wide or no-ball must be bowled again by the bowling side. This means that we could in theory have an infinite number of consecutive extras. The sum of the series is the expected total runs from non-legitimate balls for a given *i* and *j*.

Note that we make the simplifying assumption that a batsman cannot be dismissed from a no-ball or wide. This is not technically true as a batsman can be run out from either type of

delivery, stumped from a wide or otherwise dismissed by an exceptionally rare method such as handled the ball. Noting that when a wide or a no-ball is bowled *i* does not change, in the rare cases in our data set where the batsmen was dismissed from a no-ball or a wide we allocate this dismissal to the legitimate ball of the group of balls that occur at *i*. Additionally, we note that we run our dynamic programme twice, once including and once excluding the conditions variable χ as an explanatory variable in determining r_{ij}, λ_{ij} γ_{ij} and τ_{ij}.

The state space for this model consists of 3311 cells (301 possible values for *i* multiplied by the 11 possible values for *j*), of which the 311 cells involving either $i = 301$ and/or $j = 10$ are terminal cells indicating that the innings has been completed. It is very unlikely that a team could survive until ball number 300 without having lost any of its wickets. It is even more unlikely that a team could lose all ten wickets while still being on ball number one. Indeed we render this situation impossible with our simplifying assumption that a wicket cannot fall on a non-legitimate ball. However, we cover the entire state space with our estimated models. This is partly for reasons of completeness, but more importantly because the value of V in any one cell has an effect on the value of V in earlier cells.

A parametric approach to the modeling of $r_{ij}, \gamma_{ij}, \tau_{ij}$ and λ_{ij} ensures that the V-functions in the thin data cells takes into account the data in the thick data cells. We could, in theory, calculate the V-functions simply by taking average additional runs, rather than running a dynamic programme. Ideally, this approach would result in very similar results to our dynamic programme in the thick data cells; however, it would lead to unreliable information in the thin data cells and no information at all in the data cells without any observations.

We are subsequently able to calculate the value of the value function, $V(i, j)$, for each i and j, by backward induction, using Equation (19). The result of this analysis is that we have a complete set of values for expected additional runs for any possible first-innings situation.

5.6 Estimating the value function (without conditions)

In this section we consider a world in which we have no information about the ground conditions. We simply build our expected runs, probability of a wicket, probability of a wide or no-ball and expected runs from a wide or no-ball functions using information about the current state (i, j). It is important to build this model for two reasons. First, we want to see how well our model fits the observed average additional runs in each state for which the data are reasonably thick. Second, we want to assess the impact of including our newly created (in Chapter 4) ground conditions variable in the model in order to gain an understanding of the benefit of including such a variable.

5.6.1 The expected runs functions

Because there are 300 different non-terminal values of i but only ten different non-terminal values of j, it makes sense to investigate the data by wicket. Initially, we plot the average runs scored from each ball of the innings for a given number of wickets lost, j. This gives us an idea of the shape of the function that links average runs with the ball of the innings, i. For innings balls less than or equal to 90 in the non-power-play era and less than or equal to 120 in the power-play era, there are greater restrictions on the places that a fielding captain can

position his fielders.[11] We consider this a structural break in the data and we model the period in which these additional restrictions are in place separately.

Figure 5.3 is a scatter plot of average runs against innings ball for situations where a team has lost two wickets. We see an increasing trend, indicating that batsmen tend to score more quickly as the innings progresses, for the same number of wickets lost. This makes intuitive sense as the balls remaining constraint becomes a larger factor, compared to the wickets remaining constraint, the later we are in the innings. We notice that our data points have a large variance very early and very late in the innings. This is due to it being a rather rare occurrence that a team would have lost two wickets at these times and so these are thin data regions and we cannot be confident about them. In Figure 5.4, the data from Figure 5.3 is repeated only for cells for which we have at least 30 observations. This plot shows substantially greater stability in the data.

[11] According to the laws, the fielding captain could select two blocks of 30 balls some time from the 61st ball of the innings until the 300th and final ball of the innings, in which the fielding restrictions would apply. In practice, it was very rare that a fielding captain would not use these power-plays at the earliest possible opportunity; that is, balls 61 to 120.

Figure 5.3: Average runs for j = 2

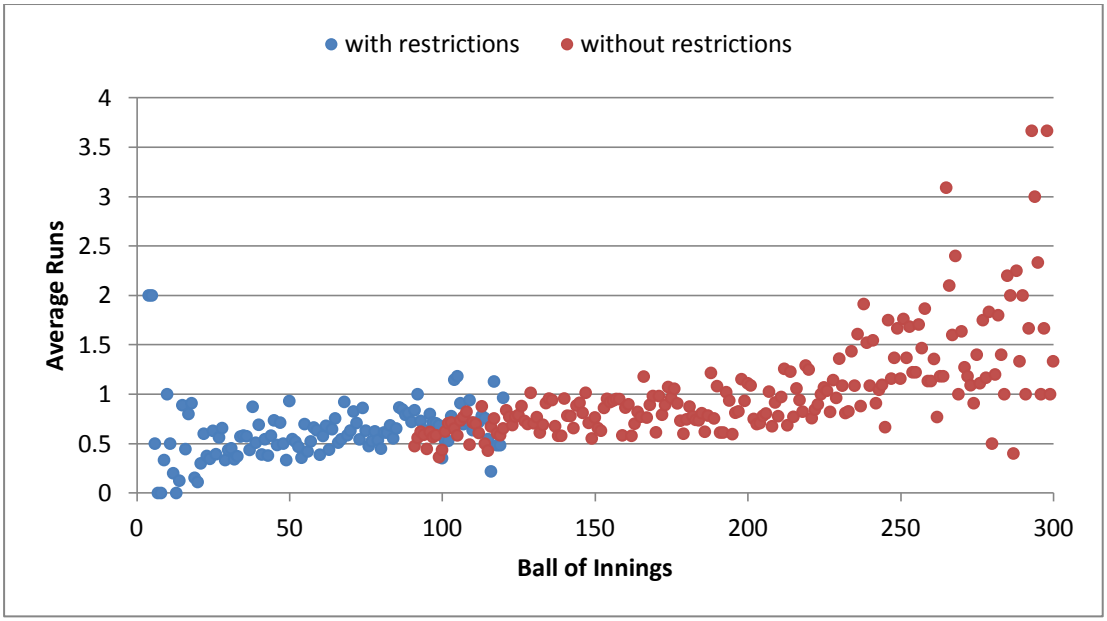
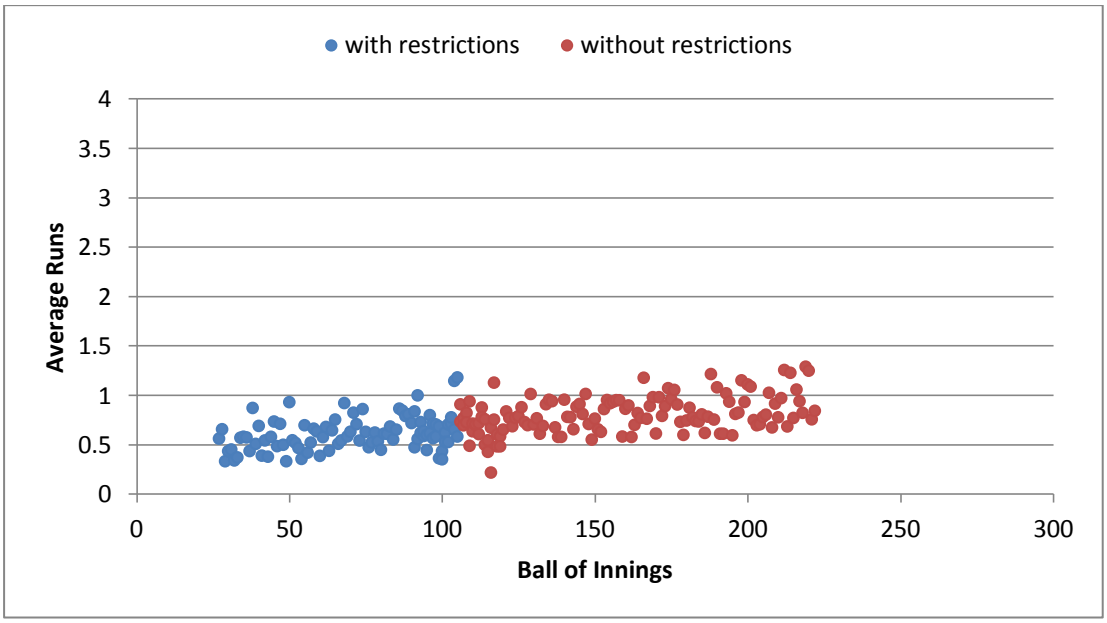


Figure 5.4: Average runs for j = 2 (thick data)



We show the average runs by innings ball within the thick data regions for $j=0$ and $j=7$ in Figures 5.5 and 5.6, respectively. Note that we only have information about the very start of the innings for $j=0$ and about the very end of the innings for $j=7$. It is clear that we

need to make some assumptions in our parametric models for dealing with the regions where data are missing.

Figure 5.5: Average runs for $j = 0$ (thick data)

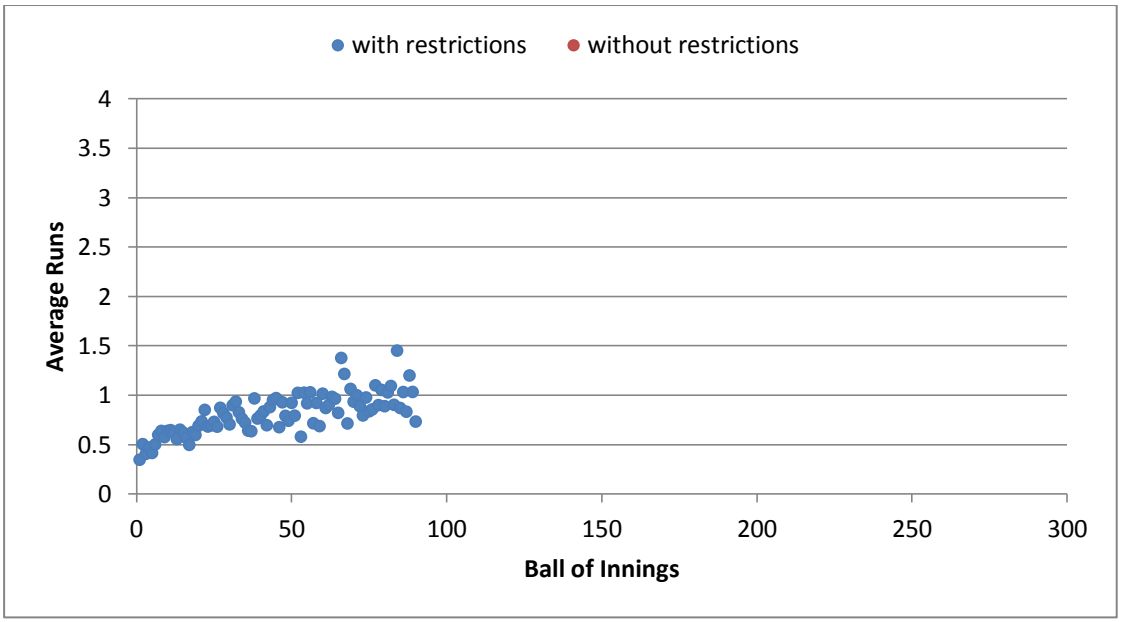
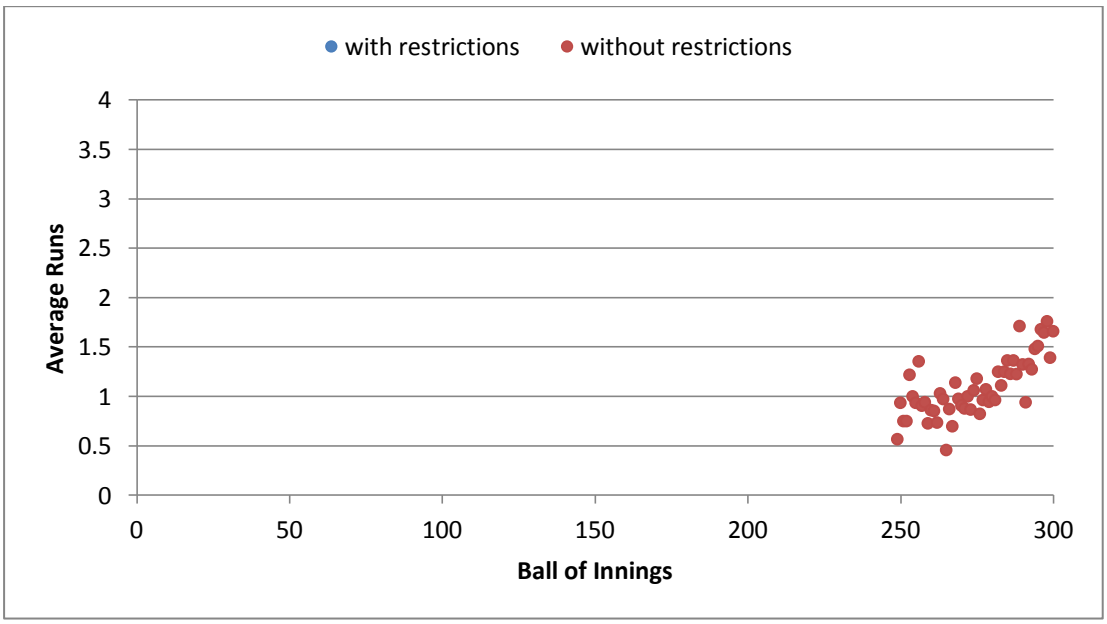


Figure 5.6: Average runs for $j = 7$ (thick data)



The value function (outlined in Equation (18)) requires $E[r_{ij}]$, the expected runs per ball in state (i, j). We have the choice between modelling this variable directly as an ordinary least squares regression with r_{ij} as a continuous dependent variable, or running an ordered Probit regression with r_{ij} as an ordinal dependent variable. The ordered Probit model gives the probability of r_{ij} taking each value and we then simply take expectations to determine the expected runs. We decide on the ordered Probit approach as this gives the curve a larger amount of flexibility. In addition, our second-innings models outlined in Chapter 7 require the probability of scoring each number of runs and choosing an ordered Probit model allows us to take a consistent approach in the models of the two innings.

We create two ordered Probit models: one for the part of the innings where the fielding restrictions apply and one for the part of the innings not affected by the fielding restrictions. Technically these should perhaps be referred to as the part with extra fielding restrictions and the part with basic fielding restrictions, as there are *some* restrictions on the field at all times; however, for simplicity we refer to the "with restrictions" period and the "without restrictions" period. The fielding restrictions can cause a sizeable positive variation in the run-scoring ability of the batsmen.

To enable us to run the regression for each stage of the innings only once, rather than running a separate regression for each value of *j*, we create dummy variables for each wicket lost. We define these dummy variables as

$$W_k \in \begin{Bmatrix} 1, where\ j = k \\ 0, where\ j \neq k \end{Bmatrix}.$$

The independent variables are now $\{i, W_0, W_1, ..., W_9\}$.

Before we perform the regressions, we need to consider our lack of data for some cells. This thin data is most apparent in situation where a team has lost many wickets after a relatively small amount of balls. In fact, there is significant data in the fielding restriction overs only for wickets zero, one, two, three and four, as shown in Table 5.2. We initially only estimate the model for these first five values of *j*. We discuss and impose a simple alternative method of creating the fielding restriction region models for the higher numbers of wickets later in this chapter.

Table 5.2: Number of observations by wicket and restrictions

Wickets Lost	With Restrictions	Without Restrictions
0	11390	903
1	9988	5285
2	6396	8329
3	2826	11734
4	813	11607
5	323	7998
6	33	5861
7	0	4260
8	0	2245
9	0	1246

The first model to consider is that using the "with restrictions" data. The criteria for accepting or rejecting a variable in the model are a combination of the significance of the variable and intuition about the game of cricket. We are primarily looking for a model that will fit the data well. The coefficients and p-values of each of our variables in our selected model are given in Appendix C. The coefficients of the dummy variables for wicket and the innings ball variable had low p-values when modelled separately; however, the interaction terms appear

to add substantial information, hence our decision to include them. The dummy variable indicating whether we are in the non-power-play or power-play era prove to have little significance and has been excluded from the model. In order to give the model some extra flexibility, we define a new parameter as

Let $i_2 = \max(0, i - 48)$

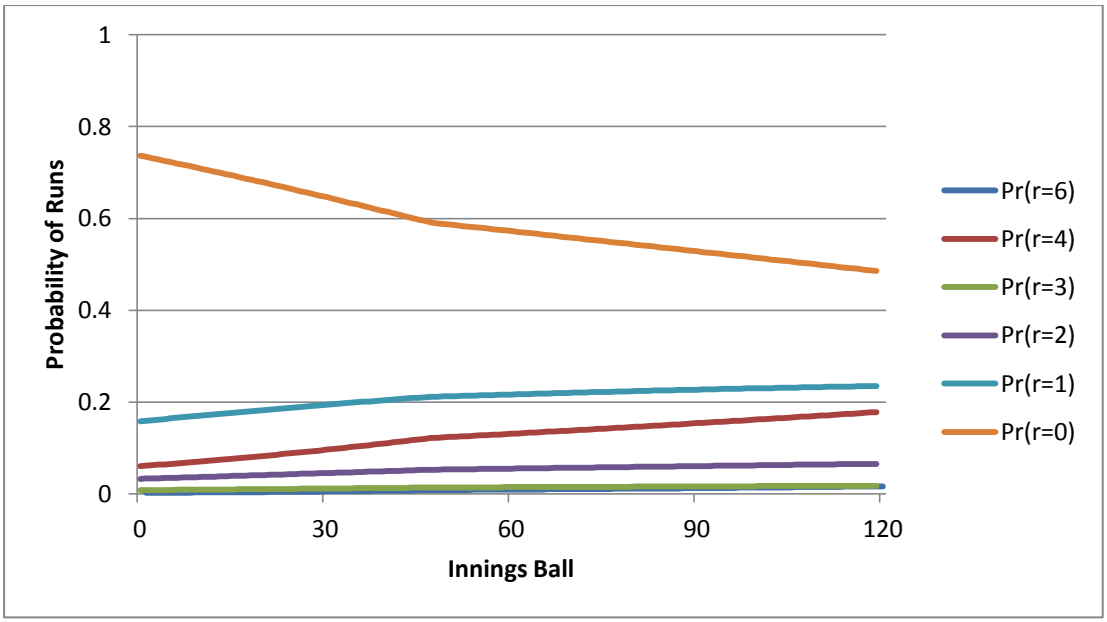
We hope that this new parameter will allow for us to take into account any effect caused by the ball being new and the teams getting used to the nature of the ground conditions, which often takes several overs. In addition, it might take into account any effect from the bowling team's two best seam bowlers being used to bowl the first overs of the innings. We choose the 49^{th} ball to change the slope as this is the start of a new over and close to the average halfway point of the "with restrictions" period, bearing in mind that this period was 90 balls for some games and 120 balls for a smaller number of games.

We note that the p-values are comparing the parameter with the base case, which is the value of j that we choose to not include a dummy variable for, in this case, $j = 4$. For our dummy variables and interaction terms involving dummy variables, we focus more on the size of the coefficient than the p-value when determining whether the variable should be included or not.

This model determines the probability of scoring each number of runs from a ball in a given state (i, j). We plot these probabilities in Figure 5.7. It is clear that dot balls (zeros) are by far the most likely outcome, but their probability decreases as the innings progresses, for a

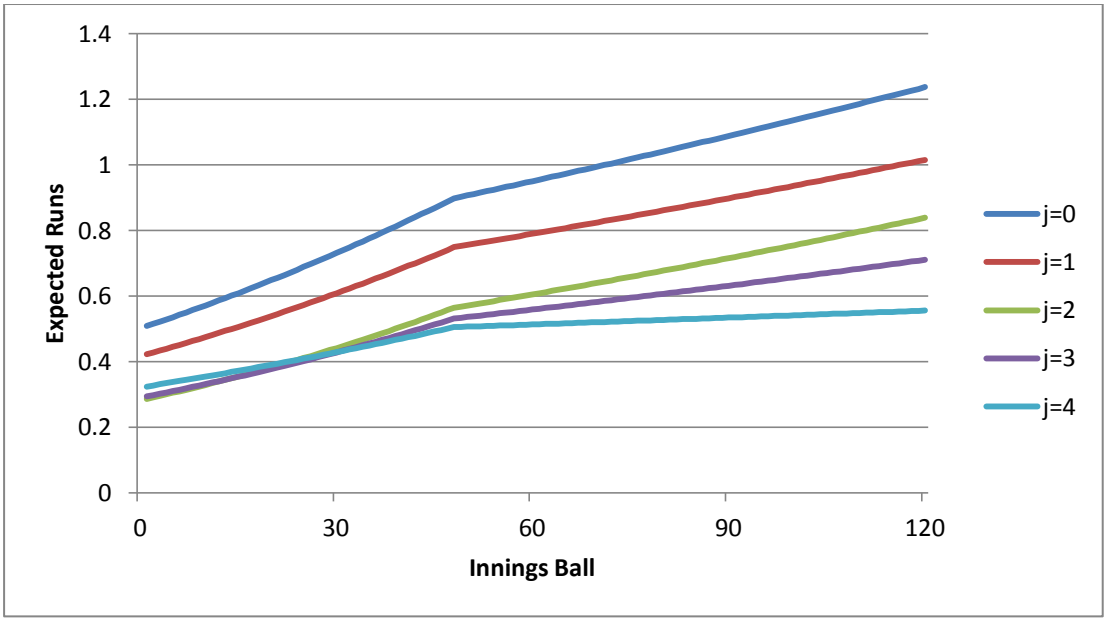
given number of wickets down. The probability of the scoring shots increases as the innings progresses.

Figure 5.7: Probabilities for *j=0*



Once we have the probability of scoring each number of runs from a given ball and wicket, we are able to take expectations to determine the expected runs for each ball and wicket. We plot these expected runs functions, for wickets zero to four, in Figure 5.8. We see two clear patterns; the scoring rate increases as the innings progresses, for a given wicket, and the fewer wickets the team has lost, the greater the rate of increase.

Figure 5.8: Expected Runs Functions - with restrictions

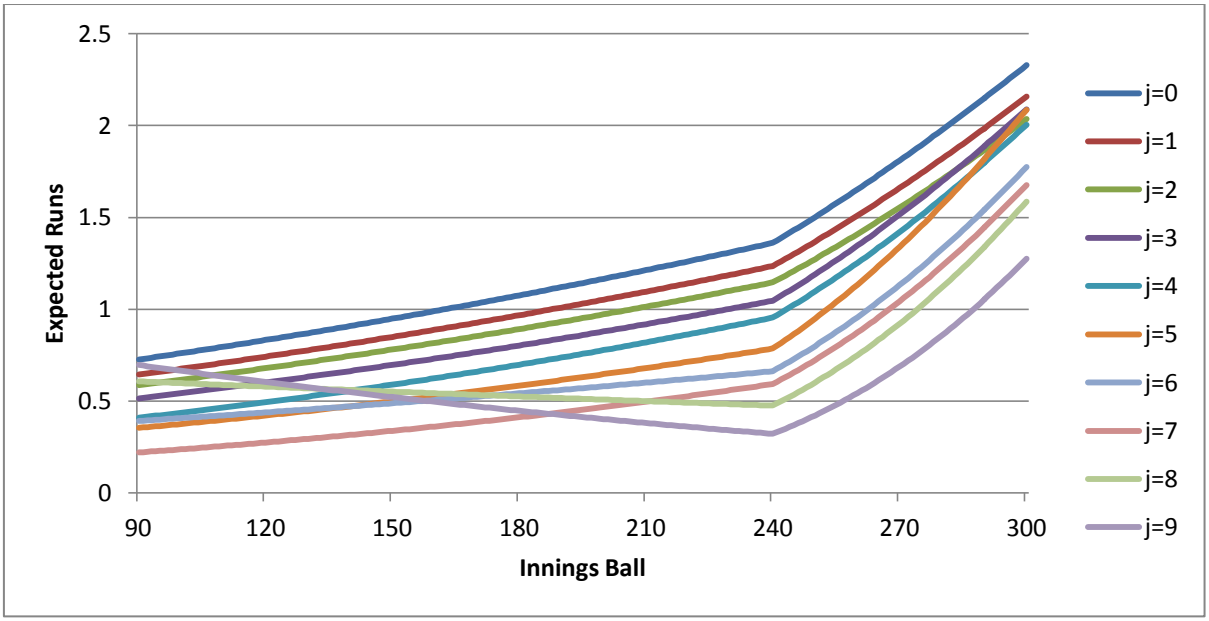


We now turn our attention to modeling the "without restrictions" data. In this data set we discover that the functional form implied by the Probit model restricts the shape of our curves by more than is appropriate for this data - therefore we create an additional variable.

Let $i_3 = \max(0, i - 240)$

The new variable allows the curves to slope steeply upward in the later balls of the innings, where the data suggest that they should. Due to limited amounts of data for some of the lower values of *j*, we assume that values of *j* less than three will all share a common slope. Note that the power-play variable (pp=1 where we are in the power-play era) makes a reasonable difference in the "without restrictions" period. The results are shown in Appendix C and the plot of the expected runs is shown in Figure 5.9.

Figure 5.9: Expected Runs Functions, without restrictions, pp=0



We have a predominantly consistent set of expected runs functions, generally upward-sloping as we get closer and closer to the end of the innings, with an acceleration of the scoring rate after our structural break at $i = 241$. The only exceptions are the curves for $j = 8$ and $j = 9$, which slope downwards until the structural break - we put this down to a lack of data early in the innings. It is simply very rare for a team to have lost a high number of wickets early in the innings and while one of the useful aspects of parametric modeling is that it provides the ability to extend the model outside the range of the data, we need to be careful that we do not overuse this to the point where it severely contradicts the things that we know about cricket. Later in this chapter we will modify this with a simplifying assumption. We also have a lack of data for the early wickets, late in the innings, so we make a simplifying assumption about these regions too. We leave these adjustments for later in the chapter as we want to use the wicket functions, to be outlined in the next section.

5.6.2 The wicket functions

The next stage of the building of the dynamic programme is to create the λ_{ij} function for the probability of losing a wicket in state (i, j). In this case we have a binary dependent variable, v_{ij}, defined as

$$v_{ij} = \begin{cases} 1, & \text{if a wicket falls in state } (i, j) \\ 0, & \text{otherwise} \end{cases}$$

We begin by plotting the distribution of average outs per ball, in order to investigate the relationship between the probability of a wicket and the ball of the innings. Note that we express the value as the average outs per ball, but we calculate this in groups of one over (a period of six balls delivered by the same bowler) here, rather than by each individual ball as we did for the expected runs functions in the previous section. This is because a wicket is a relatively rare event and there are many states (i, j) in which no wicket has fallen in our data set; therefore, we extend the period over which each average is calculated in order to obtain a smoother series.

The thick data are defined as 180 observations or more for the over group and we plot the thick data averages for $j \in \{2, 0, 7\}$ in Figures 5.10, 5.11 and 5.12, respectively. We note that there is a small upward slope, with a substantial increase in slope towards the end of the innings, noticeable in the chart where $j = 7$.

Figure 5.10: Average outs for j = 2 (thick data)

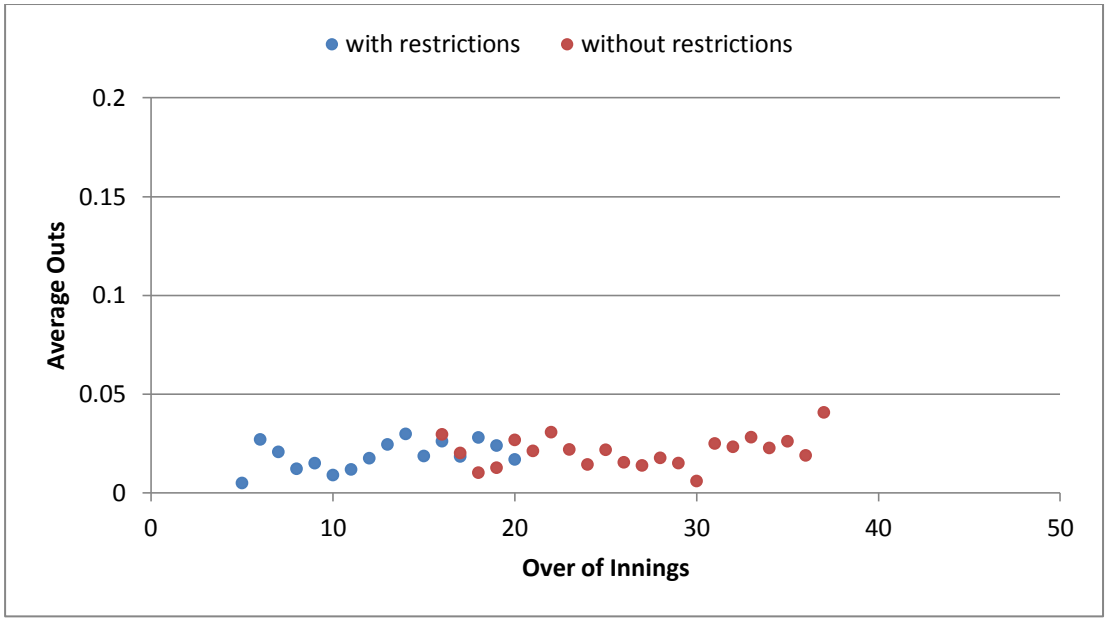


Figure 5.11: Average outs for j = 0 (thick data)

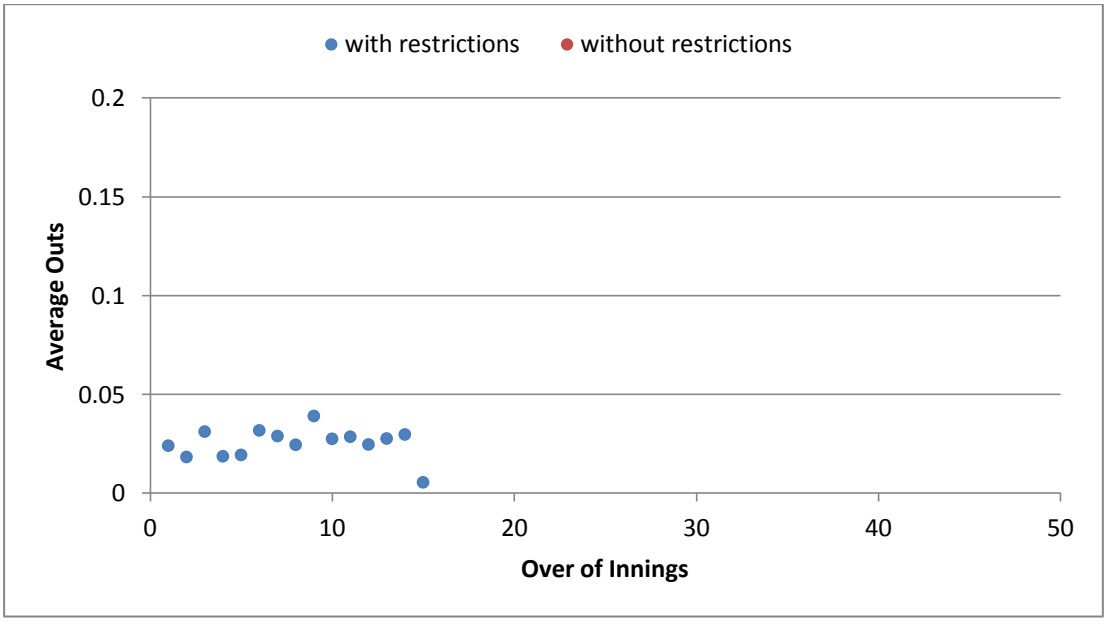
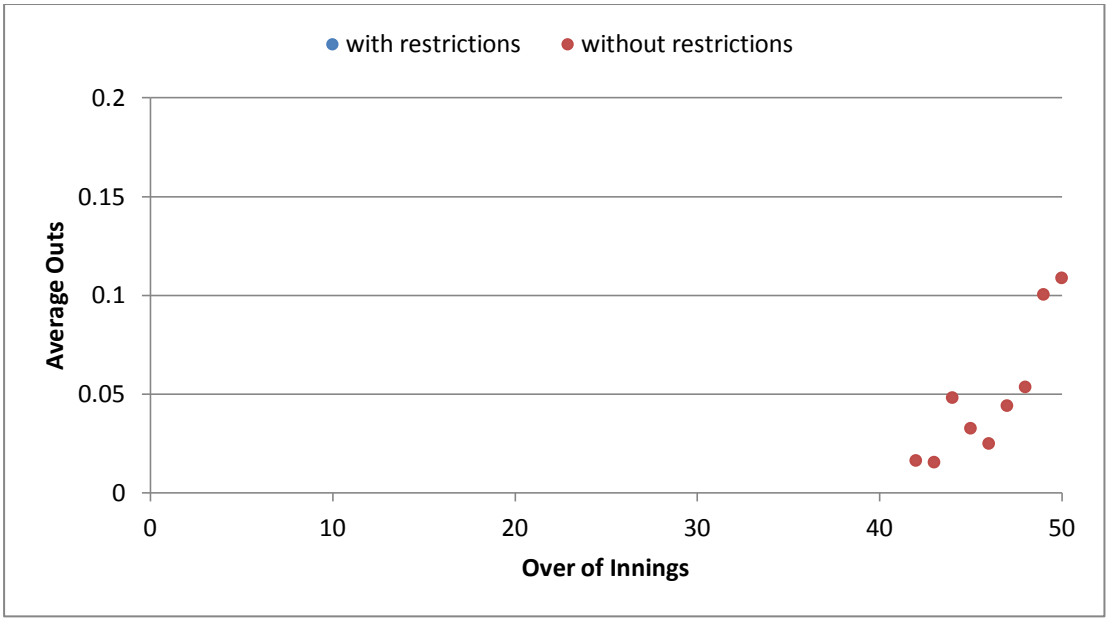
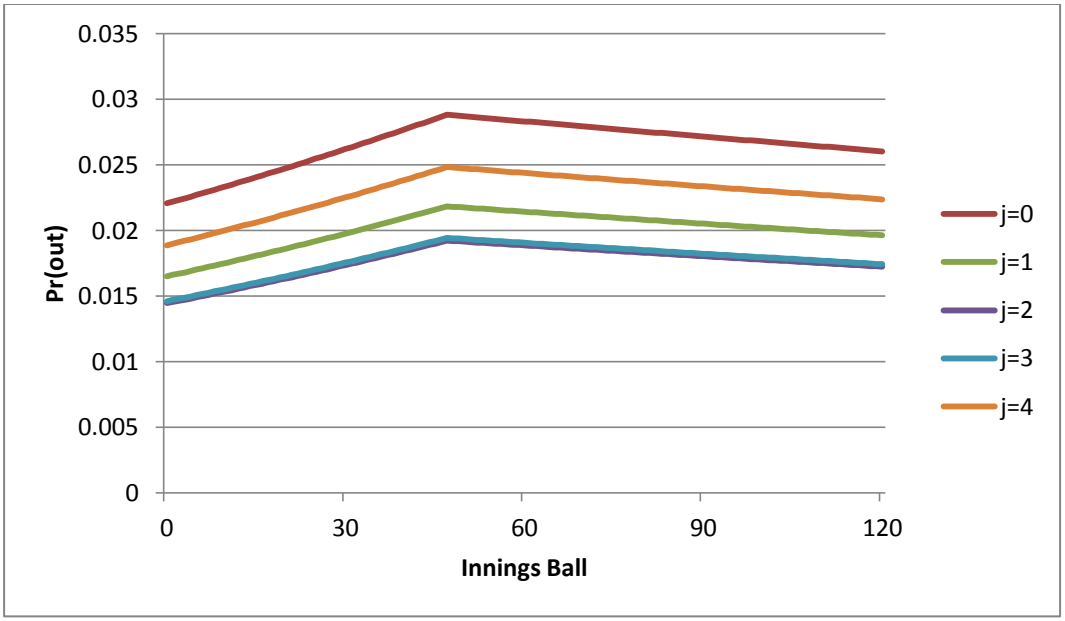


Figure 5.12: Average outs for j = 7 (thick data)



We run a Probit model on the dependent variable λ_{ij}. As with the ordered Probit for the runs functions, we run the model separately for the "with restrictions" and "without restrictions" data. It is much more difficult to fit this model than the runs model as we effectively have less information, despite the same sample size, due to the fall of a wicket being a relatively rare event. This means that we cannot find a model with coefficients and p-values that we are comfortable with by including all the interactions of i and j that we included in the runs model. The p-values are high and the coefficients do not seem consistent with our cricket intuition. We decide that it is prudent to remove these interaction terms; however, to find the balance between fitting the data well and ensuring that we do not over fit the model, we include our i_2 variable to give the model extra flexibility. The coefficients and p-values of the "with restrictions" model are shown in Appendix C and the graphical representation of these probabilities is shown in Figure 5.13.

Figure 5.13: Probability of wicket functions – with restrictions

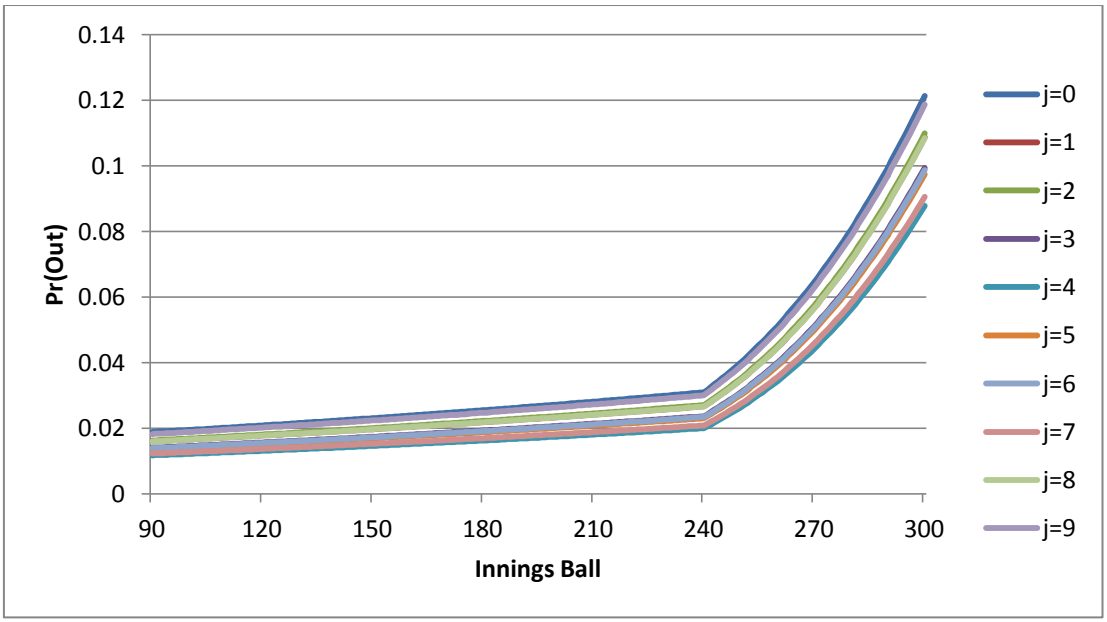


At first glance, it appears that there are only four series plotted in Figure 5.13; however, this is due to the series for $j = 2$ and $j = 3$ being almost identical. We note that the relationship between the series is not straightforward as a trade-off may exist between the higher ability of the batsmen who tend to be at the crease when a low number of wickets have fallen and the fact that, for a given value of *i*, these same batsmen can afford to take more risks as the team has a higher number of wickets in hand. We see that there is a relatively low probability of being dismissed on the first few balls of the innings, probably due to the batsmen being cautious at this time. The probability of a wicket then increases, presumably as the batsmen take increased amounts of risk in a bid to score more quickly, then starts to decrease from the structural break, which we hypothesise is due to the effects of the new ball and the better bowlers wearing off.

We now consider the "without restrictions" model. Once again, the rare nature of the event of a wicket falling leads us to take the conservative approach of excluding the interaction

terms between i and j. The parameter estimates are given in Appendix C and the probability of a wicket functions are shown in Figure 5.14.

Figure 5.14: Probability of wicket functions – without restrictions



We see that our modeling strategy has led to a very consistent set of functions, in terms of their shape. Each value of j can be grouped with other values to form the four discernible lines that appear on the graph. It is apparent that the probability of a wicket increases rapidly after our defined structural break; in this period teams are entering what is traditionally known as "the last ten over slog" or, more succinctly, "the death". We note that, as with the "with restrictions" functions, these functions are not monotonic in j, as a result of the complicated relationship between batsmen ability and acceptable risk; that is, the players higher in the order tend to be more skilful but also can reasonably take more risks, for a given i, as there are more wickets remaining.

5.6.3 Modifying the expected runs and wickets functions

In our estimation of the expected runs and the probability of a wicket functions, we have tried to find the right balance between fitting the model to available data and allowing for reliable inference about the regions where no data exists. Our goal is to estimate the expected runs and probability of a wicket for all 3000 non-terminal cells of the state space, no matter how unlikely a given state is to occur. We recognise that our fitted models will be very accurate in the thick data areas, but with a decreasing level of accuracy the further away from the thick data areas that we get. We decide to make adjustments to the extremely unlikely regions of our state space by making some assumptions.

The first problem occurs because we do not have much information about what is likely to happen if a team loses a lot of wickets early in the innings. In fact, we do not even have functions for the "with restrictions" period for wickets lost of five or more. Fortunately, some cricket knowledge can assist us with making an adjustment here. We know that a batting team faces two constraints, the number of balls remaining $(301-i)$ and the number of wickets remaining $(10-j)$. The players should adjust their risk strategy based on a combination of these two constraints. If a team has lost an unusual number of wickets for the number of balls gone in the innings, there comes a point where the batsmen recognise that there is very little chance that they will make it through the 300 balls of the innings; that is, the wickets constraint will almost certainly be the constraint that ends the innings.

We know, from our probability of a wicket functions, the probability that a team will lose a wicket at any point of the innings. It is possible to calculate, through backward induction, the probability that a team will survive the full 300 balls.

Let ϕ be the probability of the team surviving the full 300 balls

The Bellman equation is

$$\phi(i, j) = \lambda_{ij}\phi(i+1, j+1) + (1-\lambda_{ij})\phi(i+1, j) \quad (20)$$

with terminal values

$$\phi(301, j) = 1$$

$$\phi(i, 10) = 0$$

We use Equation (20) to determine the probability of the team surviving the full 300 balls from any situation (i, j). Once we have calculated these survival probabilities, we assume that players will not change the way they play based on the number of balls remaining if the survival probability is less than 0.1. We find the greatest value of i for which the survival probability is less than 0.1 and we impose the expected runs and probability of a wicket implied by that value of i on all previous values of i for that j. In the situation where the survival probability does not go below 0.1 but we reach the beginning of the "without restrictions" period and there is no function for the "with restrictions" period, we assume that the expected runs and probability of a wicket implied by the first value of i in the "without restrictions" period applies to the entire "with restrictions" period for that wicket. Note that we start this procedure by calculating the probability of survival for the entire function for $j = 9$, make the adjustment as described, then calculate the probability of survival for $j = 8$ and so forth. This

ensures that the adjusted probability of a wicket for any value of j affects the survival probability for earlier values of j.

The second problem with our runs and wickets functions is that we do not have great confidence in the models where a team has lost a very low number of wickets as we approach the end of the innings. Again, we use our knowledge of cricket to make an assumption. We know that a team will usually have a minimum of six highly skilled batsmen. On the last ball of the innings, a team will be trying to score the highest number of runs possible from that ball, without regard for the probability of getting out, except insomuch as that a high probability of getting out has an impact on the expected runs from that ball. This means that the skills and risk attitude to those batsmen at the crease where $i = 300$ and $j \in \{0,1,2,3,4\}$ should be similar; therefore, the expected runs and probability of a wicket should be similar in these five states. To make our adjustment, we assume that the cell $(i = 300, j = 4)$ gives the correct expected runs and probability of a wicket, as this is the most data-rich cell out of the five. We note that that our structural break at $i = 240$ is where the expected runs and probability of a wicket functions start sloping up sharply and we therefore fit a linear function from $(240, j)$ to $(300, 4)$ for $j \in \{0,1,2,3\}$. In this way we are using both the original function (in an area where it is more likely to be accurate) and an assumed end point from our most accurate function with two batsmen at the crease at the end of the innings.

We show the modified expected runs and probability of wicket models in Figures 5.15 and 5.16, respectively, for the example of the non-power-play era. These figures show the combined model with both the restrictions and non-restrictions period. In the non-power-play era, there is therefore a structural break at $i = 90$ for the end of the fielding restrictions. In the

power-play era, the structural break occurs at $i = 120$. We see that the higher order wickets now result in similar expected runs and probability of a wicket as we get towards the end of the innings, while the functions are flat in the regions where the ball of the innings is unlikely to affect proceedings. In particular, where a team is nine wickets down, the modified functions imply that batsmen in the last wicket partnership score rather quickly in the period where they are unlikely to see out the innings, but at an extremely high probability of dismissal. This fits well with our cricket intuition, which is that the last batsman is generally so poor that his batting partner will substantially increase the amount of risk that he takes, in order to get the most runs possible before the partnership ends. Indeed, this often causes the better batsman to be the one dismissed.

Figure 5.15: Expected runs functions, combined, pp=0

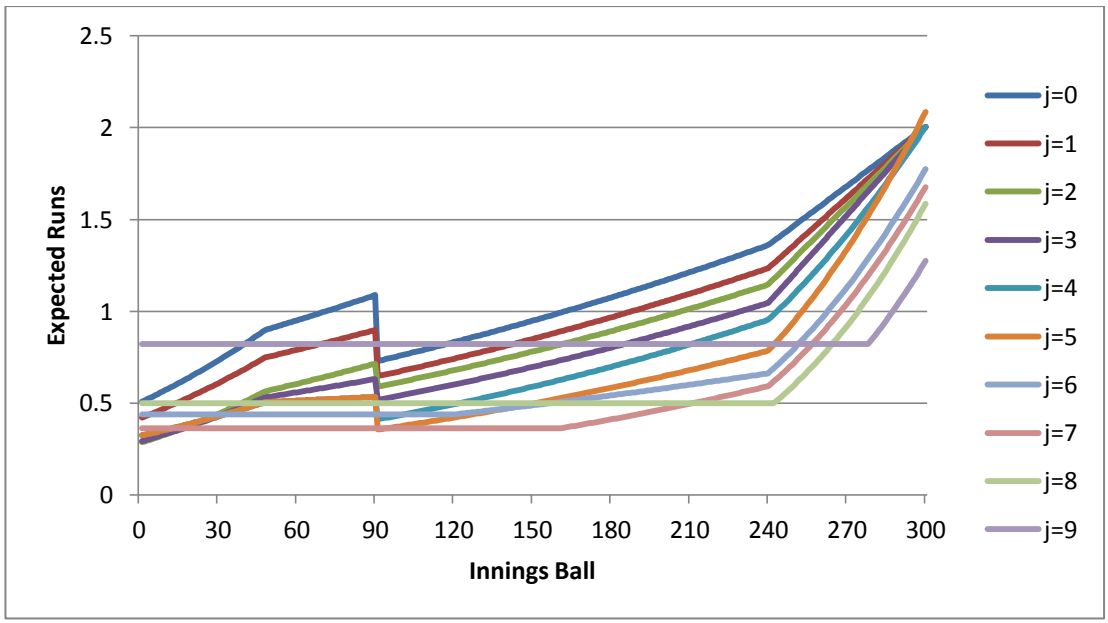
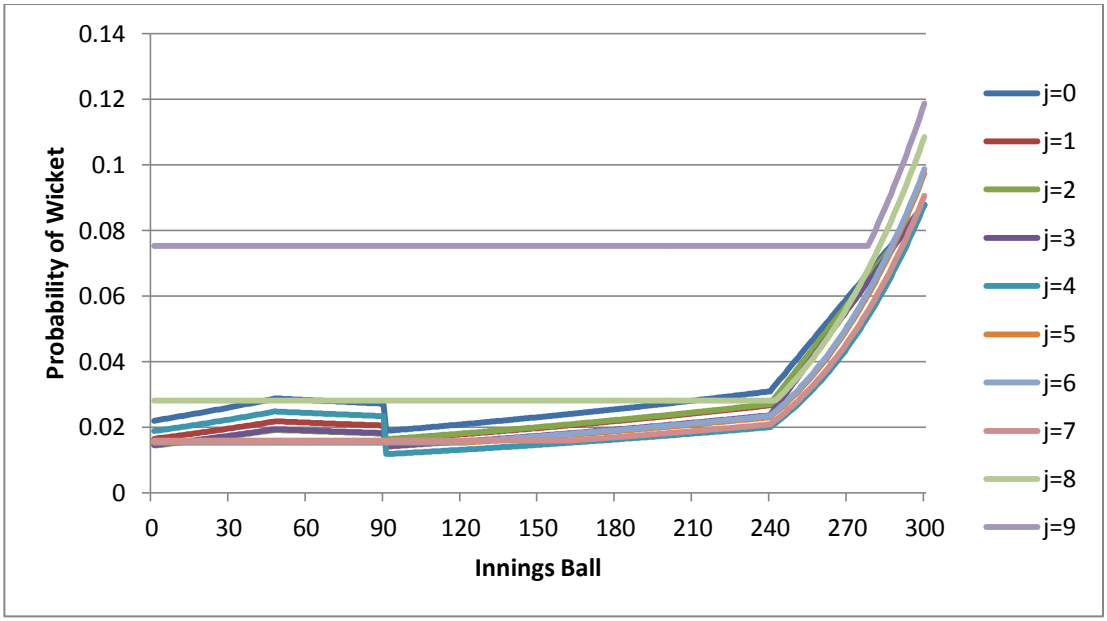


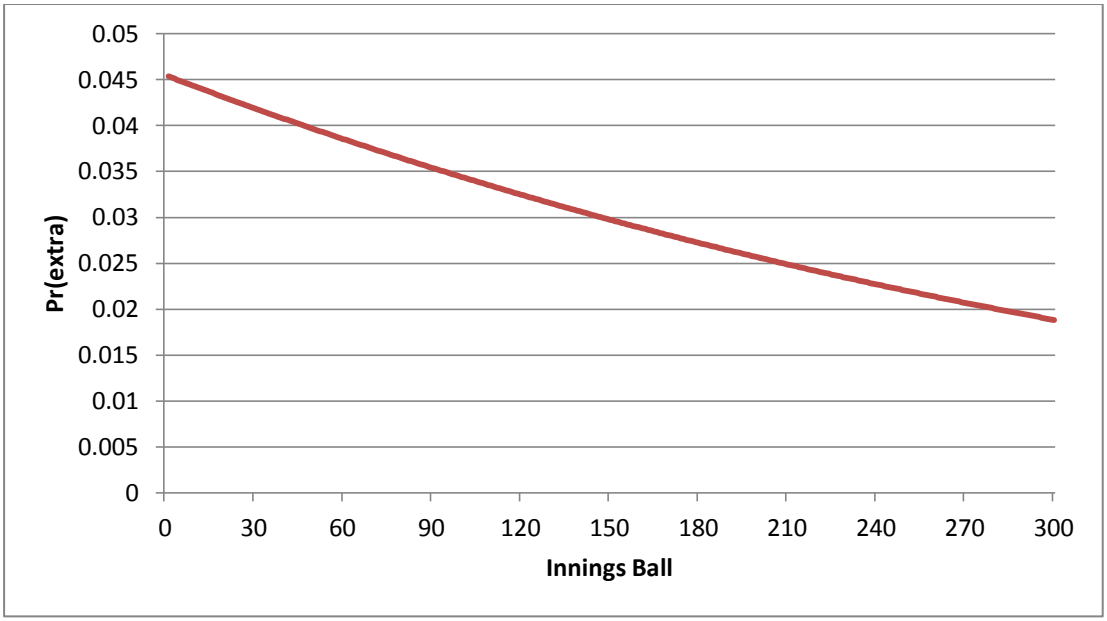
Figure 5.16: Probability of wicket functions, combined, pp=0



5.6.4 Calculating the probability of a wide or no-ball

The next variable for us to estimate is the probability that a bowler will bowl a wide or a no-ball. These result in one run to the batting team and do not count as one of the 300 balls. Since it is unlikely that the position of the fielders would have any impact on the probability of the bowling of a wide or no-ball, we create just one Probit model here, rather than separate functions for the "with restrictions" and "without restrictions" data. The regression results are given in Appendix C and the graph is shown in Figure 5.17. We note that the bowlers are most likely to bowl wides or no-balls near the start of the innings. This is likely to be due to the new ball being harder to control and the bowlers taking some time to get into the correct bowling rhythm.

Figure 5.17: Probability of a wide or no-ball function – combined model

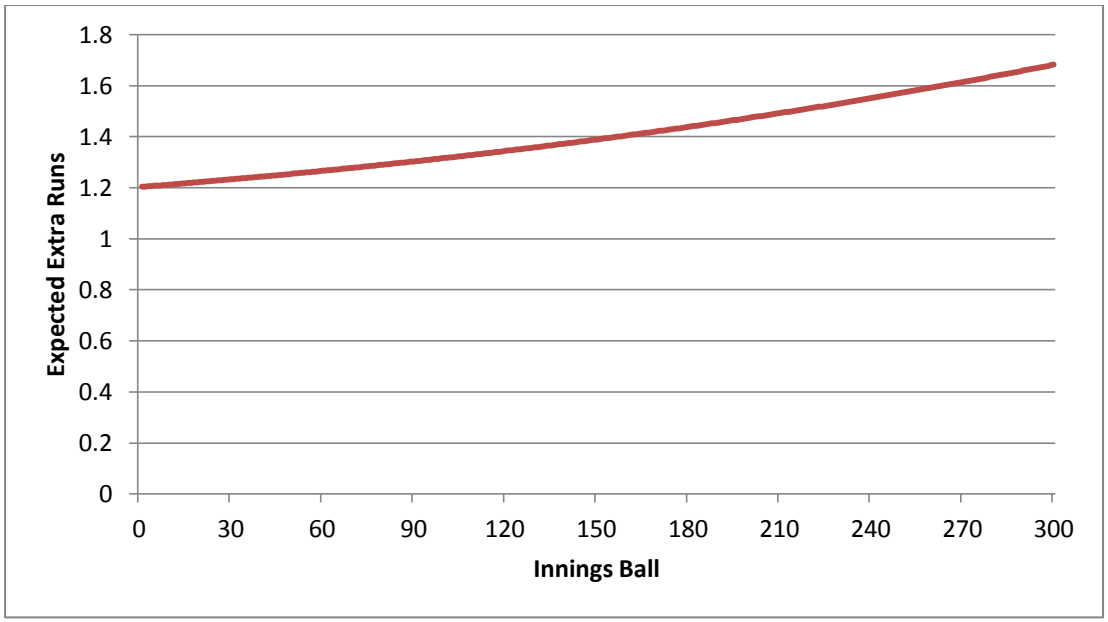


5.6.5 Calculating the expected runs from a wide or no-ball

The penalty for bowling a wide or a no-ball is one run; however, it is possible that more runs could be scored from these deliveries. In the case of a wide, the ball may be so wide that the wicket-keeper cannot stop it and the batsman may run extras or it may go all the way to the boundary for four (which is given as five wides, for the boundary plus the extra). In the case of a no-ball, the batsman may hit the ball for runs or byes or leg byes may be taken (all of which get added to the penalty run). We model the runs from a wide or no-ball as an ordered Probit model with possible values of $\tau_{ij} \in \{1,2,3,4,5,7\}$. As with our regular runs functions, we take expectations to plug into our dynamic programme. We note that there may be a small effect of the batsman being able to hit no-balls for more runs in the "with restrictions" period than the "without restrictions" period; however, we decide that this effect is fairly insignificant and we

model the runs functions from the combined data set. The parameter estimates are given in Appendix C and the graph is shown in Figure 5.18.

Figure 5.18: Expected runs from a wide or no-ball function – combined model



It is clear that bowlers will be unlikely to concede significantly more than the one run penalty if they bowl a wide or no-ball at the start of an innings; however, they will concede on average about 1.7 runs if they commit this sin at the end of the innings. This is almost certainly due to no-balls being hit for runs as it seems unlikely that wicket keepers suddenly become more likely to fail to stop wides cleanly.

5.6.6 Solving the dynamic programme

We now possess, for each state (i, j), estimates for $E[r_{ij}]$, λ_{ij}, γ_{ij} and τ_{ij}. This is all that is required to solve our value function, $V(i, j)$, by backward induction. Recall Equation (19).

$$V(i,j) = E[r_{ij}] + \lambda_{ij} V(i+1, j+1) + (1-\lambda_{ij}) V(i+1, j) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

We plot the resulting V-functions, from the non-power-play era, in Figure 5.19. In Figure 5.20 we add the power-play-era models as a dashed line for comparison to the non-power-play-era models. Recall that the only differences in the two models are a dummy variable for the era, which was used only in the runs functions in the "without restrictions" regression, and the fact that the "with restrictions" period runs for the first 120 balls, rather than 90, in the power-play era. Not surprisingly, the biggest difference occurs around the point of the difference in rules, for the early wickets.

Figure 5.19: V-functions, non-power-play era

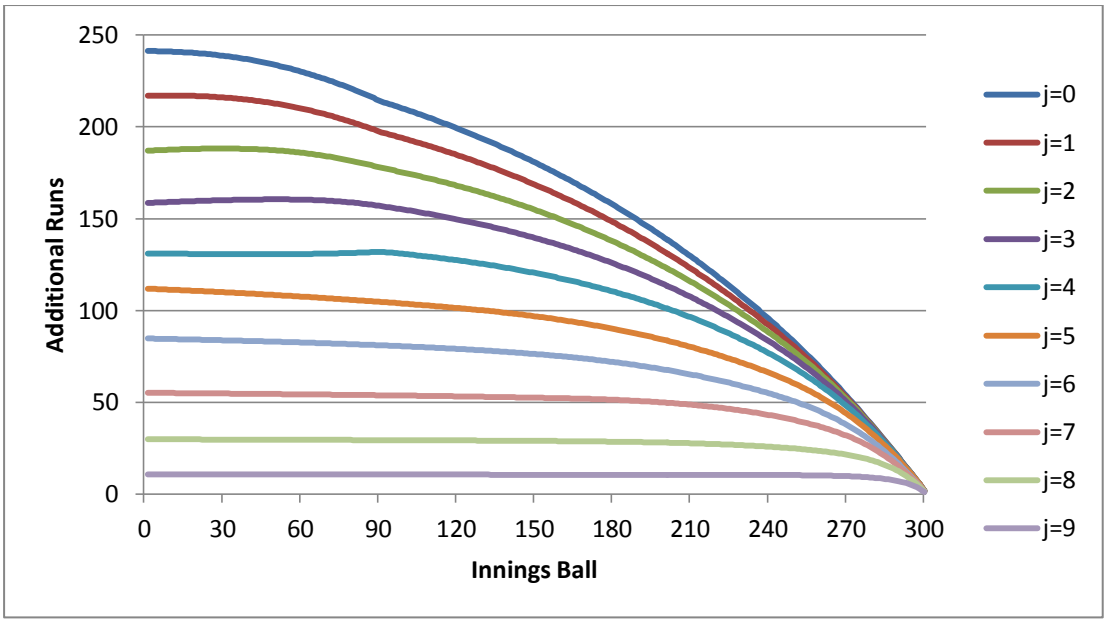
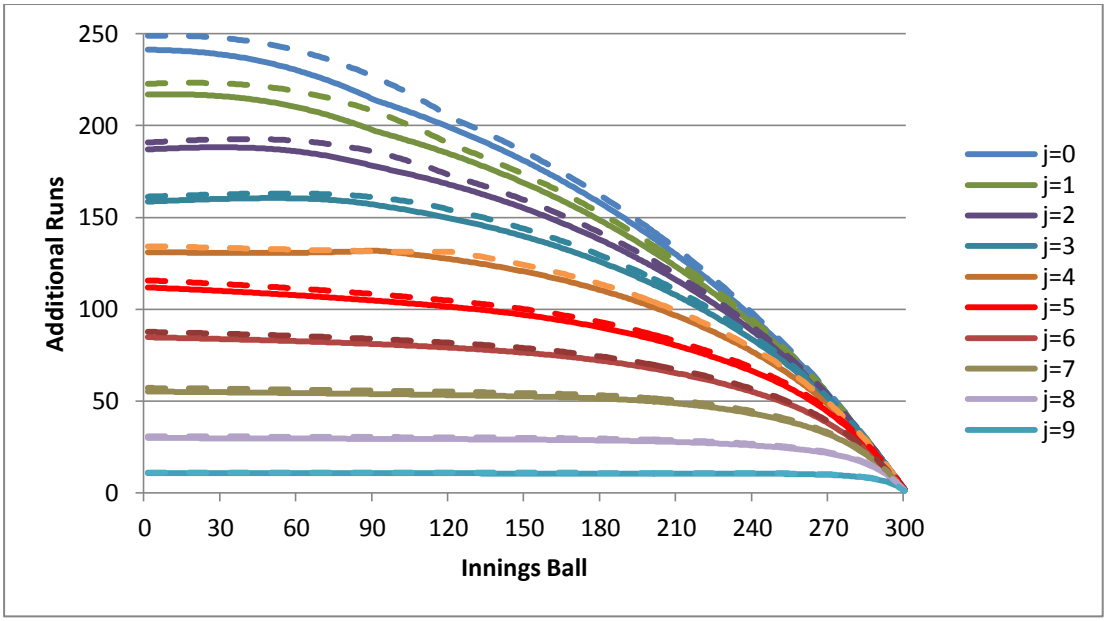
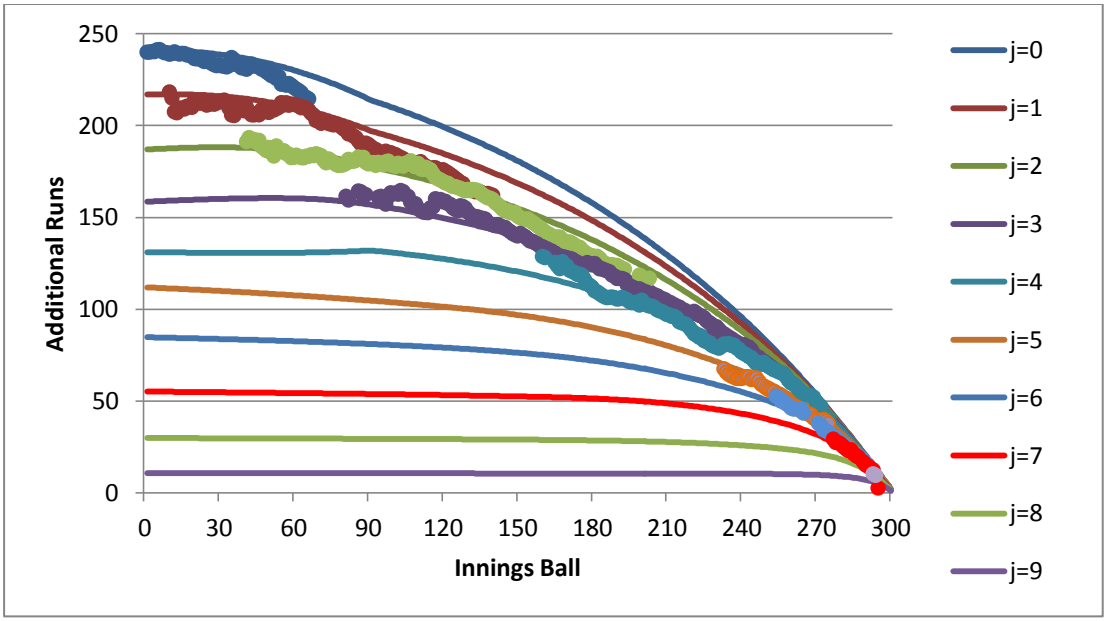


Figure 5.20: V-functions (dashed line indicates power-play-era)



Our expected additional runs functions predict future scoring from any possible first-innings situation, no matter whether the situation is common or rare. A simple average of the additional runs should also be a reasonable estimate where our data are thick. In Figure 5.21 we compare our V-functions with the average additional runs scored for all situations in the non-power-play era where we have at least 30 data points. We overlay our V-functions with solid dots of the same colour, to represent the average additional runs.

Figure 5.21: V-functions versus average additional runs, non-power-play era



It is clear that our V-functions fit the thick data well, aside from over-predicting the actual average additional runs achieved when one wicket is down in the middle stages of the innings. We note that our functions provide an indication of what should be possible, given average performance. One explanation for the discrepancy is that it is possible that teams who have made a good start to their innings by only losing one wicket at the end of the fielding restriction period, have then generally gone on to perform worse than average during the rest of the innings. The power-play-era data are a similarly good fit, albeit with fewer thick data areas due to the smaller sample size. In Table 5.3 we calculate a measure of the fit of the model to the thick data as the root weighted mean square error. We define this as

$$RWMSE = \sqrt{(V_{ij} - Average_{ij})^2 \left(\frac{count_{ij}}{count_j} \right)}$$

where

$Average_{ij}$ is the average observed additional runs in that state

$count_{ij}$ is the number of observations in that cell

$count_j$ is the number of observations for that value of j.

Note that we are measuring the distance between the average additional runs and the modeled expected additional runs, rather than measuring the distance from the expected additional runs of the actual additional runs from each individual match. We note that our worst fit, for $j = 1$ in the non-power-play era, is different from the data on average by only about 3% of the average first-innings total score.

Table 5.3: Weighted Mean Square Error of Fit

j	Non-power-play	Power-play
0	4.292	4.863
1	7.225	5.925
2	4.371	6.833
3	4.263	4.370
4	3.391	1.492
5	3.531	2.656
6	1.751	N/A
7	1.074	2.295
8	0.664	N/A
9	N/A	N/A

5.7 Estimating the value function (with conditions)

In Chapter 4, we calculated a variable, χ, to represent the ground conditions occurring in each match. We inferred this variable using only information about the first-innings scores and results of matches and we created a conditional distribution for ground conditions χ, for each first-innings score and result of the match. These conditional distributions were almost exactly normal; therefore, we assume normality and our information set about the conditions in each game consists of a mean and a variance.

There are two choices of ground conditions model: the version where we used a single data set of 784 matches, or the version where we split the data set into matches involving an additional five overs of fielding restrictions (the power-play era). The former was slightly better at predicting first-innings scores, given draws of conditions from the conditional distributions; however, since we want to create our V functions for the two separate eras, we decide to use the latter. In order to include the ground conditions variable in our Probit regression models, we use the following procedure.

1. Determine the first-innings score and result of each match.
2. Apply the mean and variance of the conditional distribution of χ, as implied by the analysis in Chapter 4, to each match.
3. For each ball, simulate a random number, θ, from the standard normal distribution.
4. The simulated value for ground conditions is $\chi = \theta\sqrt{\sigma^2_{\chi|S,\omega}} + \mu_{\chi|S,\omega}$.
5. Repeat steps 3. and 4. n times to generate n values of χ for each ball.
6. In the regression models, give each observation a weight of $\frac{1}{n}$.

The above procedure means that we have effectively multiplied the size of our data set by n. Each observation is repeated n times in the data set with the ground conditions variable being the only variable that is changing with each repetition. The weight variable has no impact on the coefficients of the Probit regressions, as each observation has the same weight; however, it ensures that our p-values are still meaningful as we tell SAS about the weight when we run the regressions. The multiplication of the data set without weighting would result in incorrect (very small) standard errors and p-values.

We are being slightly inconsistent by taking this approach. In Chapter 4, one of our necessary assumptions for being able to determine the conditional distribution of χ was that χ does not change during a game. Theoretically, this implies that we should simulate one value of χ, apply it to all balls in that game, and repeat that process n times, rather than generating n values for each ball. The reason for our approach is that we have limited computing power and we therefore cannot simulate as many values as we would choose to with unlimited computing power. By simulating just n values for each game, we risk distorting the data for an entire game if our simulation procedure results in a set of values of χ not representative of the underlying conditional distribution. On the other hand, each individual ball has only a very minor role in determining the coefficients of the Probit models and we therefore are able to get a better representation of the conditional distribution of χ by simulating values for each ball. We note that the simulation process takes slightly longer using our approach, but the size of the resulting data sets are identical, for a given value of n; therefore, there is no carry-over difference in the time taken after the multiplication of the data set is complete.

The choice of n is a balance between accuracy and the time available, given our computing power. Clearly, the higher n is, the more accurate our models will be. In order to determine the most appropriate n, we run our entire dynamic programme multiple times for selected values of n in order to determine how high n needs to be to ensure accuracy. Eventually, we settle on $n=100$, meaning that our data set is 100 times its original size of approximately 91000 observations.

5.7.1 The regression equations

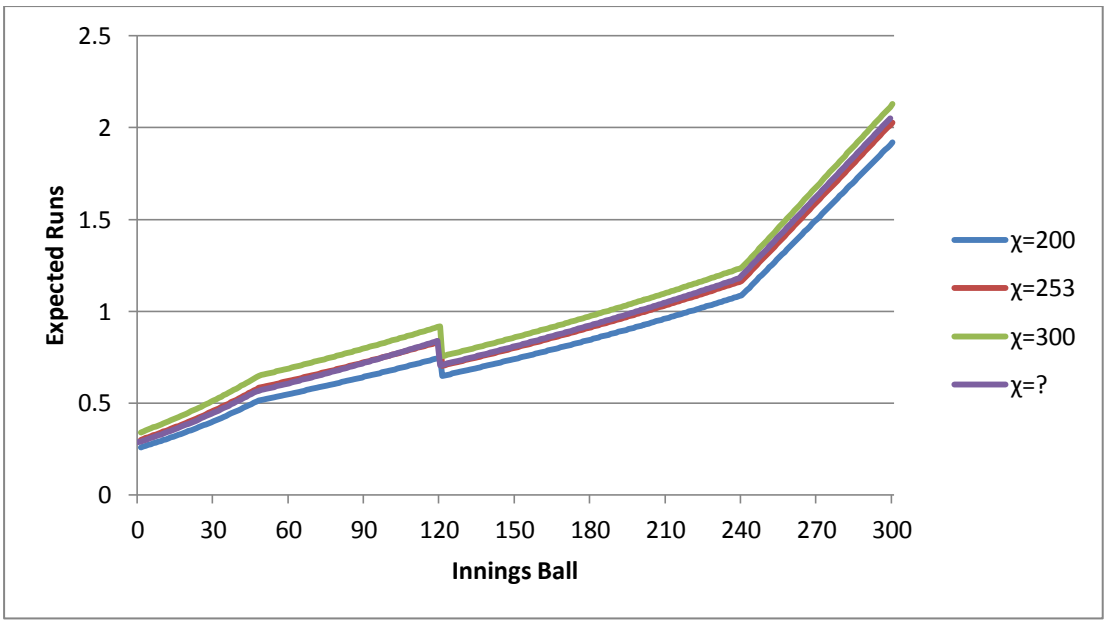
We include our simulated values for χ in our Probit and Ordered Probit models for r, λ, γ and τ. It is important to also consider the interaction of χ and i as ground conditions (particularly weather) may have a larger impact when the ball is new. We also reconsider the role of the power-play variable; this is because our estimation strategy for χ involved splitting the data into the two eras and it is possible that the inclusion of χ may change our decision to include or exclude the power-play variable in each model.

In the runs models, the only new variable included is the χ variable, as the interaction of χ and i and the power-play variable made very little difference. The variable coefficients and p-values are included in Appendix C.

Figure 5.22 displays a plot of the expected runs functions implied by the models, where $j=2$ and $pp=1$, for three selected examples of conditions along with the model created without any information about conditions (this is represented by the line labeled $\chi=?$). We deliberately choose $\chi=253$ as one of our examples as this is the overall mean first-innings

total in the power-play era. This shows that the function where conditions are average is not the same as the no-conditions function. The latter is below the average conditions function early in the innings and above it later in the innings. This is because the no-conditions model implicitly assumes that conditions are likely to be bad for batting when a team has lost two early wickets and good for batting when a team makes it a long way through their innings for the loss of only two wickets. Note that we show the models after having applied the same thin data adjustment techniques used in the unknown-conditions model

Figure 5.22: Expected Runs Functions, power-play era, *j=2*

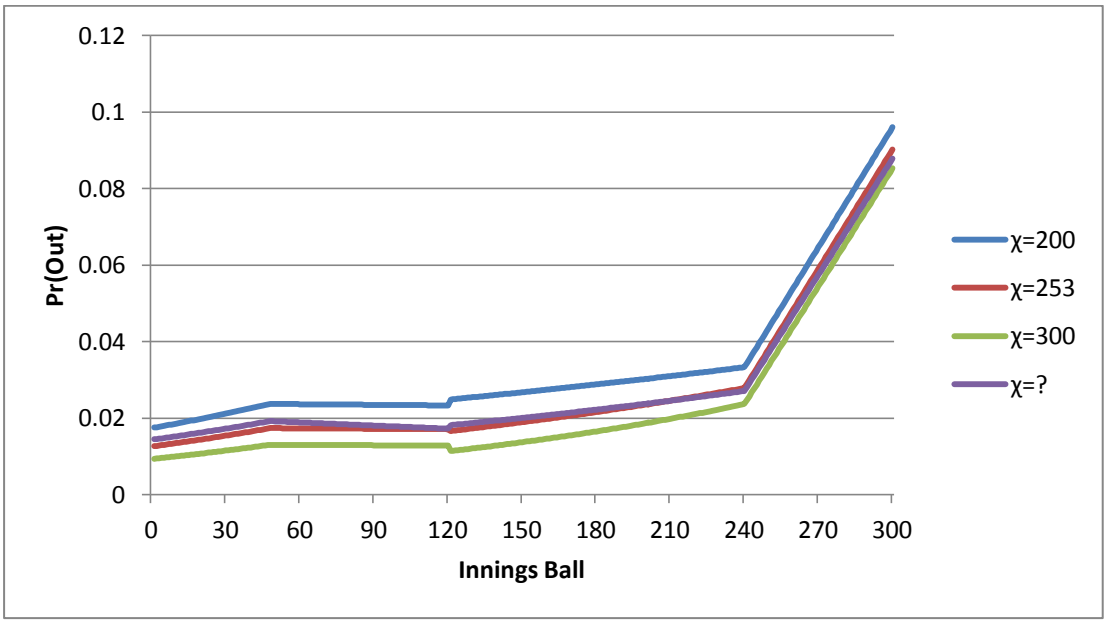


In the "with restrictions" period, the χ variable is the only new inclusion and in the "without restrictions" period we find that the interaction of χ and i is significant; therefore, it is also included in the model. The coefficients and p-values are given in Appendix C. The

positive coefficient, combined with the negative coefficient of χ, implies that the conditions have a smaller impact on the probability of a wicket as the innings progresses.

Figure 5.23 contains a graphical representation of the wicket functions. Note that again we see a difference between the model that assumes average conditions and the model that has no knowledge of conditions. This time, the probability of a wicket is higher if two have fallen early as the no-conditions model implicitly assumes a high likelihood of difficult batting conditions. The inclusion of the interaction term between ground conditions and innings ball means that the difference between the curves for the various conditions is larger at the start of the "with restrictions" period than at the end of the innings.

Figure 5.23: Wicket Functions, power-play era, *j=2*



We next add our conditions variable to the probability of a wide or no-ball model. Somewhat surprisingly, it suggests that teams are more likely to bowl wides or no-balls in good

batting conditions. We suggest that a possible reason for this is that bowlers faced with difficult bowling conditions (good batting conditions) may be willing to sacrifice some accuracy for additional pace or variation, as taking wickets is difficult in such conditions. As we expect, teams score a higher number of runs on average from wides and no-balls in good batting conditions than in poor batting conditions, presumably because it is easier to score from no-balls, which can be hit. The regression coefficients and p-values of these two models are given in Appendix C.

We show the graph of γ, the probability of a wide or no-ball and τ, the expected runs from a wide or no-ball, in Figures 5.24 and 5.25, respectively.

Figure 5.24: Probability of Extra Functions, power-play era, *j=2*

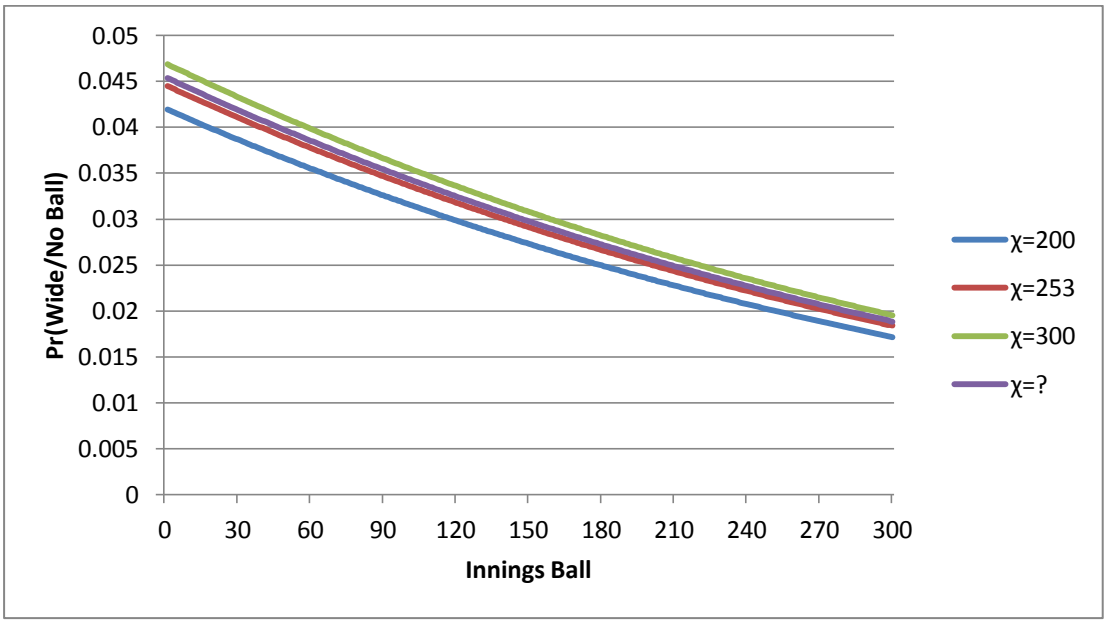
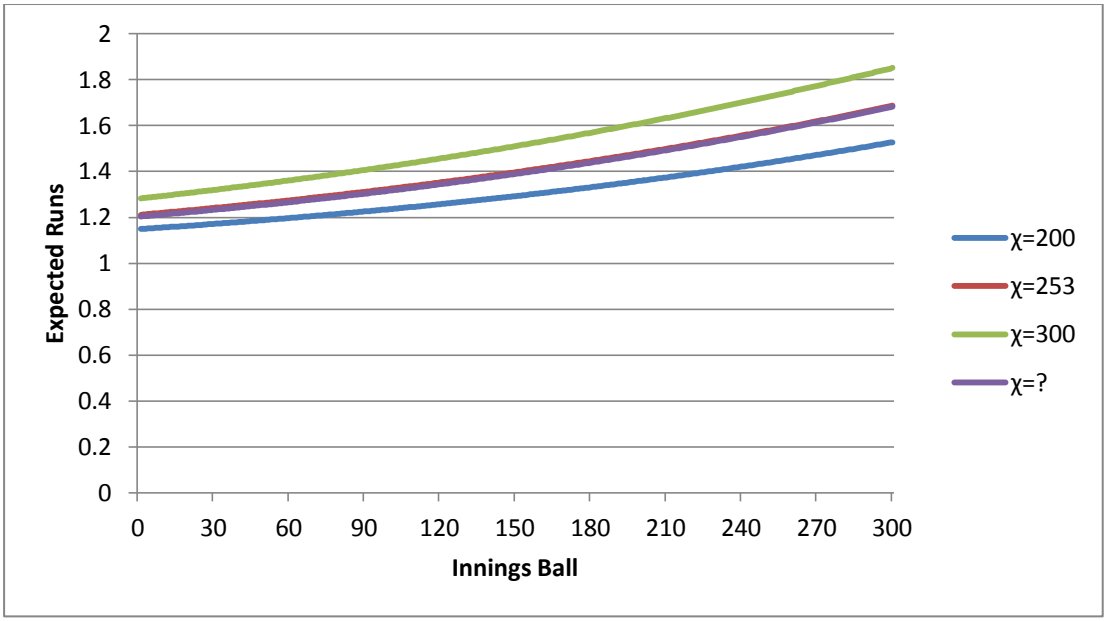


Figure 5.25: Expected Runs from Extras Functions, power-play era, *j=2*



We calculate the V-functions by solving the dynamic programme by backward induction separately for each integer value of χ in the range $[0,500]$. Define the value function as

$$V(i,j,\chi=\mathrm{X}) = E[r_{ij\chi}] + (\lambda_{ij\chi}V(i+1,j+1,\chi=\mathrm{X}) + (1-\lambda_{ij\chi})V(i+1,j,\chi=\mathrm{X}) + \frac{\gamma_{ij\chi}\tau_{ij\chi}}{1-\gamma_{ij\chi}}$$

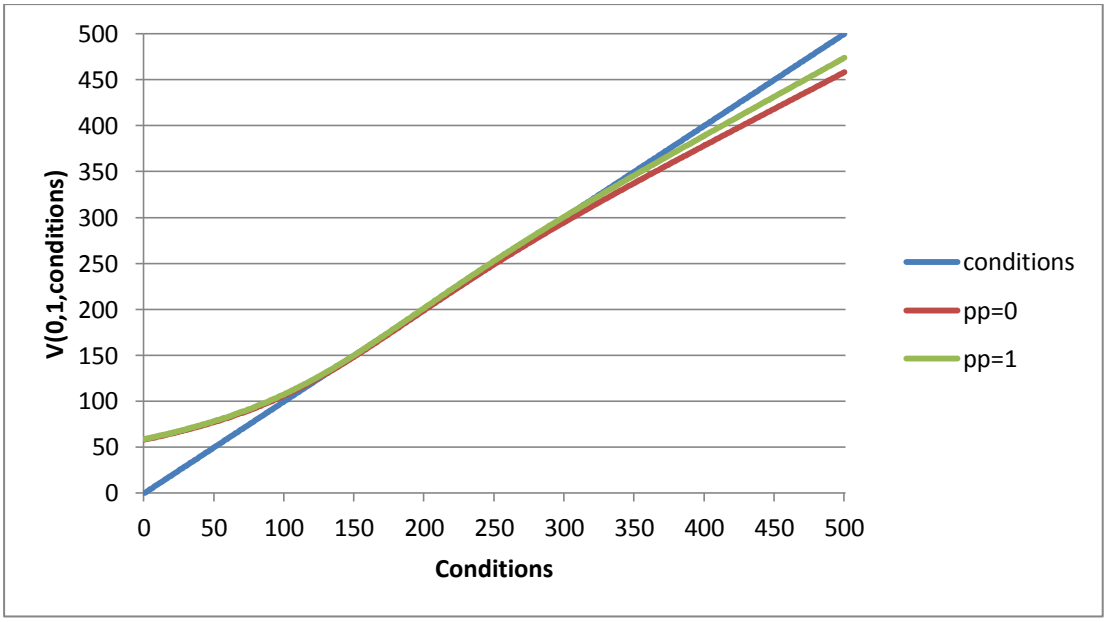
If our model is perfectly calibrated, we expect that the expected additional runs at the start of the first inning would be equal to the value for ground conditions; that is

$$V(1,0,\chi=\mathrm{X}) = \mathrm{X} \qquad (21)$$

For each era, we plot the calculated $V(1,0,\chi)$ value for each χ between 0 and 500, inclusive, in Figure 5.26. The 45° line indicates the locus of points where Equation (21) holds.

It is clear that our fit is generally very good in the range of conditions that we are likely to observe, although the model is not coping as well with the less-likely conditions. We note that these results are in line with expectations. It is not possible to observe a score of less than zero; therefore, the average score for conditions of zero is clearly going to be substantially greater than zero. Our theoretical model allows for both score and conditions to take any value; therefore, at these extreme values, it is not going to match what is observed in practice, which drives the dynamic programme. A similar argument can be made for the higher values of conditions. While it is theoretically possible to observe any score, in our data set the highest score is 434 and it is therefore no surprise that the dynamic programme finds it more difficult to fit the model for values of conditions where the lack of higher scores would mean that we would expect a substantial number of draws of conditions to be above the highest observed score.

Figure 5.26: Expected Additional Runs at start of innings versus Conditions



We test our fit to the theoretical value using the following procedure.

1. For each era, simulate 10000 values of χ using the normal distributions estimated in Chapter 4. In the non-power-play era, $\chi \sim N(238.1, 643.5)$ while in the power-play era, $\chi \sim N(250.0, 988.6)$.
2. Calculate the difference between the simulated value of χ and the $V(1, 0, \chi)$ implied by that value.
3. Square the differences, sum them, divide by 10000 and then take the square root of the result to determine the root mean square error.

Table 5.4: RMSE for fit of $V(1, 0, \chi)$ to χ

	Non-power-play	**Power-play**
RMSE	1.35 runs	2.54 runs

The root mean square errors are given in Table 5.4. It is clear that the V-functions are a good fit to the theoretical values implied by the ground conditions variable. The fit is not perfect, however, particularly at very high and very low values of χ. As we use these V-functions in determining the PPFs for batsmen in a later section of this chapter, we apply a correction to the V-functions at this point. Define the adjusted V-function as

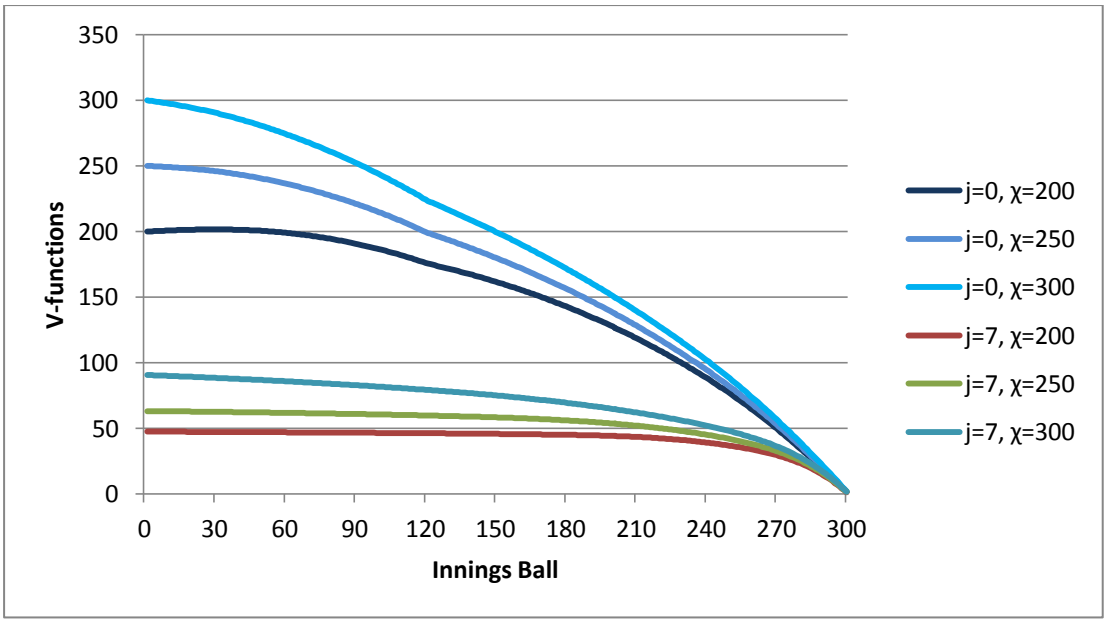
$$V_{adj}(i, j, \chi) = \frac{\chi}{V(1, 0, \chi)} V(i, j, \chi) \tag{22}$$

Equation (22) implies that all V-functions for a given *i, j and* χ will be scaled by the same factor that equates their V-function in the first cell (the start of the innings) with the theoretical ground conditions.

Now that we have our final, adjusted V-functions, it is useful to plot some examples in order to investigate the impact of the χ variable. Figure 5.27 shows the power-play-era

functions for ground conditions of 200, 250 and 300 and wickets zero and seven. These examples reveal some interesting information. First, the graph for $V(i,0,200)$ is very flat at the beginning of the innings, indicating that simply surviving and not scoring any runs is good enough to leave your expected final total unchanged, due to the difficult batting conditions. Second, early in the innings the difference between playing in 300 conditions and 250 conditions is much greater than the difference between 250 conditions and 200 conditions, for seven wickets down. This implies that, if the top order were to be dismissed early despite excellent batting conditions, the lower order could be expected to do a reasonable job.

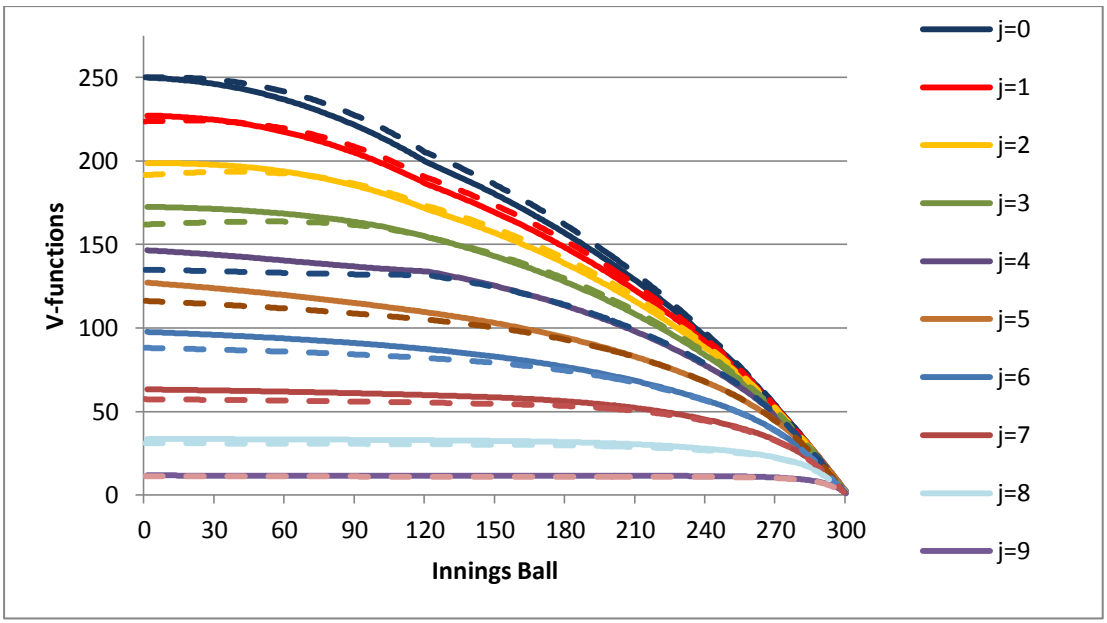
Figure 5.27: Selected V-functions



In order to once again illustrate the difference between playing in average conditions and playing in unknown conditions, in Figure 5.28 we plot, for every wicket, the V-functions for 250 conditions and for unknown conditions in the power-play era. The unknown-conditions

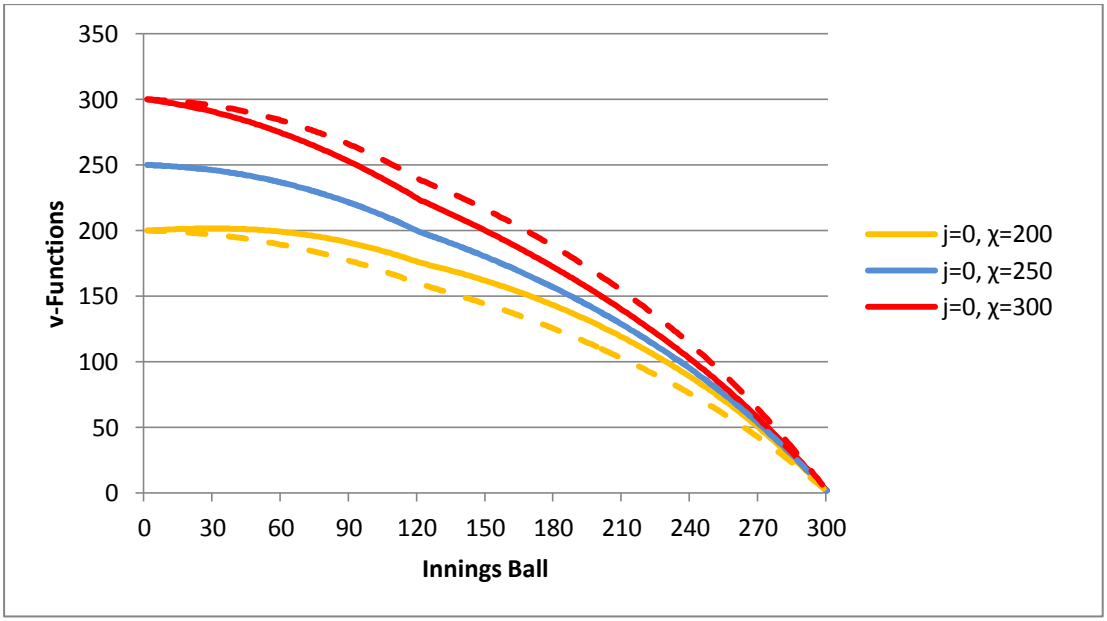
functions appear as the dashed lines. Note that, for reasons of fairness, we scale the unknown-conditions function to its theoretical initial value of 250, as we did for the V-functions with conditions. It is clear that the unknown-conditions model implicitly makes assumptions about conditions based on what has happened so far in the innings. The dashed lines are flatter than the solid lines, meaning that when wickets are lost early, the unknown-conditions model assumes that our future scoring will be lower than the average-conditions model, as it is likely that we are in poor batting conditions. By contrast, when the team gets a significant way through their innings without losing many wickets, the unknown-conditions model expects more future runs than the average-conditions model, as the former assumes the team is playing in very good batting conditions.

Figure 5.28: Average Conditions (solid lines) vs Unknown Conditions (dashed lines)



It is possible to also illustrate the benefit of solving the dynamic programme for each value of ground conditions, rather than solving it for a single value and scaling the resulting function. In Figure 5.29 we present the dynamically-solved V-functions for conditions of 200, 250 and 300 (the solid lines) and compare them to functions obtained by solving the dynamic programme only for conditions of 250 and scaling this function. These functions are for the power-play era, where $j = 0$. We see that the difference is significant; therefore, it is worthwhile spending the computer time solving for each individual value, although solving for several values and interpolating for the in-between values would provide a good approximation.

Figure 5.29: Conditions vs Scaled Conditions



5.8 Concluding remarks

The focus of this chapter has been on analysing the first innings. We have created a dynamic programme to answer questions surrounding the reasonable future expectations of a batting team from any given current position. In Chapter 6 we show how we can use the models in this chapter to construct a new variable, the cost of a wicket, which has important implications for strategy. Using the cost of a wicket as a proxy for the risk taken by batsmen we are able to estimate PPFs for individual players.

CHAPTER 6

Estimating Production Possibility Frontiers for batsmen

6.1 Introduction

In ODI cricket, a batsman faces a trade-off between the rate at which he can score runs and his probability of survival. If a batsman attempts to score at a faster rate, he is usually required to take a higher level of risk, compromising his probability of survival. Some examples include the batsman attempting to loft the ball over the fielders (risking being caught), attempting to run with a lower degree of certainty that he will make his ground (risking being run out) and attempting to hit the ball harder (risking several dismissal methods due to having less control of the bat). In this chapter, we outline a method with which we can estimate the trade-off between scoring rate and survival rate for an individual batsman. A cricket team cannot determine its optimal batting strategy without knowing the capabilities of its 11 members.

6.1.1 A hypothetical case

Our goal is to observe the trade-off between expected runs and the probability of survival. Unfortunately we cannot observe these variables directly. We are only able to observe the outcome of each ball in terms of number of runs scored and whether or not a wicket fell. To properly observe this trade-off, we require information about the risk intentions of batsmen

when particular results are achieved. We show the usefulness of this by way of a hypothetical example. The greatest batsman in the history of the game, Sir Donald Bradman, never played an ODI, as the shorter format of the game was yet to be invented. For the purpose of illustrating an example, consider for a moment the idea that Bradman had an ODI career of several years. Assume that he faced 7000 balls with the outcomes displayed in Table 6.1. Over this period, our hypothetical D. Bradman scored 6996 runs from 7000 balls faced, a scoring rate of 99.94 runs per hundred balls. Bradman was also out 70 times over this period, giving him a survival rate per ball of 99.00%. This gives us some overall idea of our hypothetical Bradman's ability but tells us nothing about how his scoring rate and survival probability change when he adopts different risk strategies.

Table 6.1: Summary of D. Bradman's batting outcomes over observed sample.

Outcome	Number of Occurrences	Percentage of Occurrences
Zero Runs	2359	46.56%
One Run	2341	33.44%
Two Runs	350	5.00%
Three Runs	175	2.50%
Four Runs	700	10.00%
Five Runs	0	0.00%
Six Runs	105	1.50%
Out	70	1.00%

We now add some more information to our hypothetical model. Imagine we knew that Bradman only ever played with two strategies, a relatively defensive strategy which we call strategy *a* and a relatively aggressive strategy which we call strategy *b*. Our hypothetical Bradman is equipped with a powerful memory and he informs us that he played exactly half the

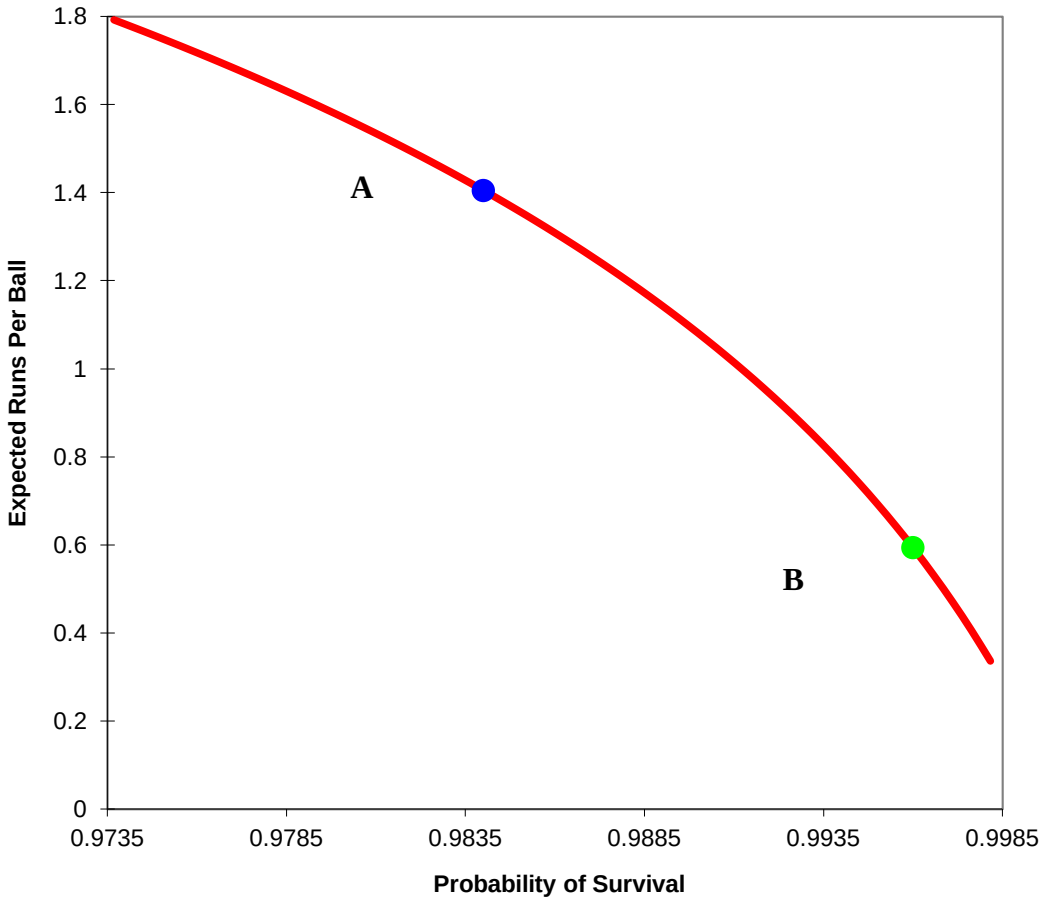
balls in his career using each strategy and is able to recall exactly which balls were played under which strategy. This information is summarised in Table 6.2.

Table 6.2: Summary of D. Bradman's batting outcomes

	Strategy *a*		**Strategy *b***	
Outcome	**Number of Occurrences**	**Percentage of Balls Faced**	**Number of Occurrences**	**Percentage of Balls Faced**
Zero Runs	2100	60.00%	1159	33.11%
One Run	1050	30.00%	1291	36.89%
Two Runs	140	4.00%	210	6.00%
Three Runs	35	1.00%	140	4.00%
Four Runs	161	4.60%	539	15.40%
Five Runs	0	0.00%	0	0.00%
Six Runs	0	0.00%	105	3.00%
Out	14	0.40%	56	1.60%

When playing strategy *a*, our hypothetical batsman D. Bradman scored a total of 2079 runs from 3500 balls, a scoring rate of 0.5940 runs per ball, while his survival rate was 99.60%. When playing strategy *b*¸ he scores a total of 4917 runs from 3500 balls, a scoring rate of 1.4048 runs per ball, while his survival rate was 98.40%. We now have two points where we know that our hypothetical Bradman was playing different risk strategies and we can infer a basic PPF. Assuming convexity of the production set, that is to say that the higher is the probability of survival, the higher is the marginal cost in terms of expected scoring rate of an additional unit of survival probability, a possible PPF for D. Bradman, if he chose to employ a full range of strategies, is displayed in Figure 6.1. Point "A" represents Bradman's aggressive strategy *a*, while point "B" represents his defensive strategy *b.*

Figure 6.1: A possible PPF for hypothetical batsman D. Bradman



6.1.2 Inferring the intentions of a batsman

In Section 6.1.1 we outlined a very simple method of finding pairs of scoring rates and survival rates to determine points on the PPF of a hypothetical batsman. In practice, this method meets with a rather large hurdle. It relies on the batsman telling us how much risk he intended to take with each ball. This information is not available to us in any practical way. Even if a batsman wanted to give us this information after a game, the chances of him remembering his exact risk intentions for every single ball of his innings are extremely slim.

It would be reasonable to assume that a batsman would take the same level of risk, on average, every time he is in exactly the same situation. Given there are 3000 combinations of i and j that make up the state space, it is infeasible to treat each state as a separate situation as, for an individual batsman, we would have a very small set of observations for each cell.

We show in this chapter that the key determinant of first-innings risk strategy should be the number of runs that a team's expected score falls by if a wicket is lost, the "cost of a wicket". The higher is this number, the larger is the potential cost to the batting team of a risky strategy and the more defensive is the optimal strategy for the current batsman to employ. There are two important things to note about the cost of a wicket. First, different states (i, j) produce very similar costs, meaning that we can collapse our 3000 different states into a much smaller number of groups, increasing the sample size in each group. More importantly, the cost of a wicket converts each state (i, j) into a cardinal variable, which means that we can use cost of a wicket to create a flexible semi-parametric model to make predictions for all cells (i, j).

A batsman's ability determines the trade-off that he faces between scoring rate and survival rate. There also exists a preference trade-off, the marginal rate of substitution (MRS), given by some utility function $U\left(E[r],\ \eta\right)$ where $E[r]$ is the expected runs from a particular ball and η is the probability of surviving that ball. We show that the MRS is equal to the cost of a wicket. By estimating a batsman's $E[r]$ and η as functions of the cost of a wicket, we are first able to identify his PPF and second able to test his strategic nous by comparing the points of his PPF at which he chooses to operate with the optimal strategy, for various values of the cost of a wicket.

6.2 Choosing the level of risk to optimise the value function

In Chapter 5 we estimated the value function, $V(i, j)$, by looking at actual behaviour. This was sufficient for predicting the expected additional runs from any state (i, j). In this chapter we are concerned with the chosen strategy of a batsman; therefore, we note that our expected runs and probability of a wicket functions do not arise automatically from each combination of *i* and *j*. Rather, they arise in part because of the strategic choice of a batsman.

Let κ = The level of aggression chosen by the batsman. $\kappa \in [0,1]$

We redefine our expected runs function as $E[r_{ij} \mid \kappa]$. To enable us to model a batsman's strategic choice in the conventional economic framework of substituting between two desirable goods, we use the probability of not losing a wicket on a legitimate ball, rather than the probability of a wicket. We define the probability of survival as $\eta_{ij}(\kappa)$.[12] When a batsman selects a strategy, he chooses a risk level κ; however, we do not know the relationship between κ and $E[r_{ij}]$ or η_{ij}. Define $\mathbb{Z}$ as the locus of points $\left(E[r_{ij}], \eta_{ij}\right)$ that define a batsman's PPF so that

$$(E[r_{ij}], \eta_{ij}) \in \mathbb{Z}$$

[12] $E[r_{ij}]$ and η_{ij} are batsman-specific, as well as depending on *i* and *j*; however, we suppress the reference to the batsman in the notation as we do not use this notation to compare different batsmen.

We assume that when a batsman is selecting a strategy, he is implicitly choosing a point on his PPF - that is, some combination of $E\left[r_{ij}\right]$ and η_{ij} where $(E\left[r_{ij}\right], \eta_{ij}) \in \mathbb{Z}$. The Bellman Equation is now

$$V^*(i, j) = \underset{(E\left[r_{ij}\right], \eta_{ij}) \in \mathbb{Z}}{Max} V(i, j)$$

where

$$V(i, j) = E[r_{ij}] + (1-\eta_{ij})V^*(i+1, j+1) + \eta_{ij}V^*(i+1, j) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

A batsman is indifferent between the set of points where $V(i, j) = K$, for any constant K.

$$V(i, j) - K = 0$$

so

$$E[r_{ij}] + (1-\eta_{ij})V^*(i+1, j+1) + \eta_{ij}V^*(i+1, j) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}} - K = 0.$$

From the implicit function theorem, then

$$\Rightarrow \quad \frac{\partial E\left[r_{ij}\right]}{\partial \eta_{ij}} = V^*(i+1, j+1) - V^*(i+1, j)$$

Let $C(i, j)$ be the cost to the batting team of losing a wicket on ball i, given they have already lost j wickets, $C(i, j) \in \mathbb{R}$.[13] so that

$$\frac{\partial E\left[r_{ij} \right]}{\partial \eta_{ij}} = -C\left(i, j\right) \tag{23}$$

Equation (23) reveals that, in any state of the first innings, a batsman's marginal rate of substitution between scoring rate and survival rate is equal to the negative of the cost of a wicket in that state. As the cost of a wicket only depends on i and j, in any state (i, j) the batsman's indifference curve is linear with slope $-C\left(i, j\right)$.

We have outlined the fact that there are different combinations of runs and survival between which the batting team is indifferent. Now we address the capabilities of an individual batsman. On every ball, a batsman must decide how much risk he wants to take. Each level of risk results in some number of expected runs and some probability of survival, for that individual batsman. These numbers can be used to form the PPF for that batsman. A batsman is optimising if he takes the level of risk that places him at the point where his PPF is tangential to the indifference curve of the team, given by the cost of a wicket at that stage of the game. That is, where the marginal rate of substitution is equal to the marginal rate of transformation.

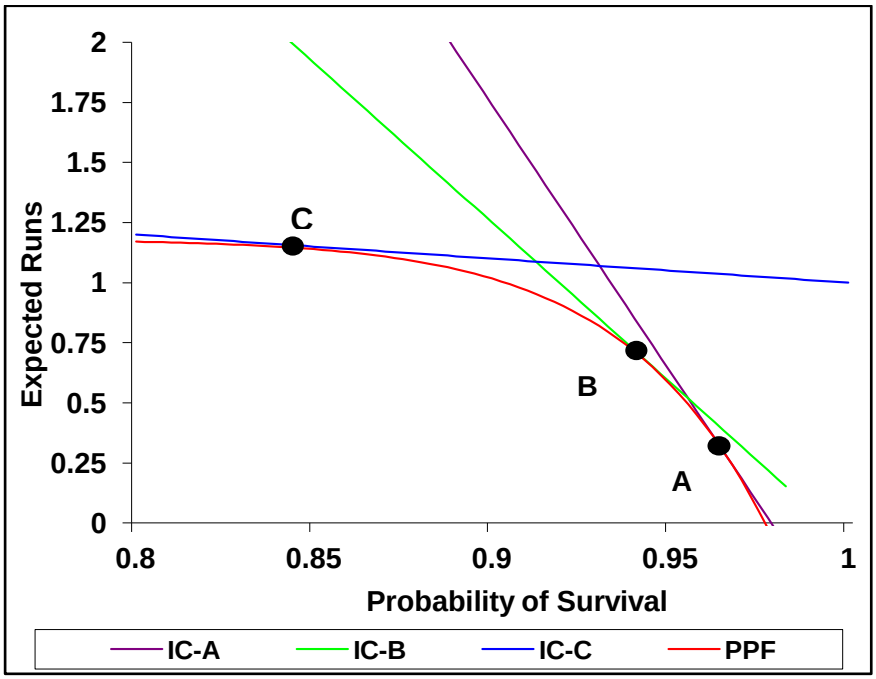
We expect that a batsman's PPF is continuous, monotonic in η_{ij} and weakly concave. This weak-concavity expectation means that a batsman will have to give up a higher amount of

[13] We note that $C(i, j)$ can technically take any real-numbered value. Situations do exist in the game of cricket where the batting team may actually be better off by the loss of a wicket, for example, if the situation demands fast scoring and the batsman at the crease is either having a bad day or simply does not have the ability to score as quickly as required. If a team is following the optimal strategy then the cost of a wicket should never be negative as the batsman has the option of retiring; however, a poorly-performing batsman's pride may well get in the way of the optimal strategy, opening the door to negative values of $C(i, j)$.

expected runs in order to get an extra unit of survival, the higher the probability of survival he already has and *vice versa*. This can be theoretically justified by the reasoning that a batsman, over the course of several balls, can reach any point on the straight line between two points of his PPF by mixing between those two strategies.

We illustrate an example of a batsman's PPF and his optimal choices in Figure 6.2, using three potential game situations. When the cost of a wicket is high, this particular batsman wants to be operating at point A, where he is taking a low level of risk and the highest attainable indifference curve is IC-A. When the cost of a wicket is a middling value, this batsman should optimally move to higher risk point B and he should move to very high risk point C where the cost of a wicket is very low.

Figure 6.2: The PPF with optimal points



6.3 Estimating the Production Possibility Frontiers

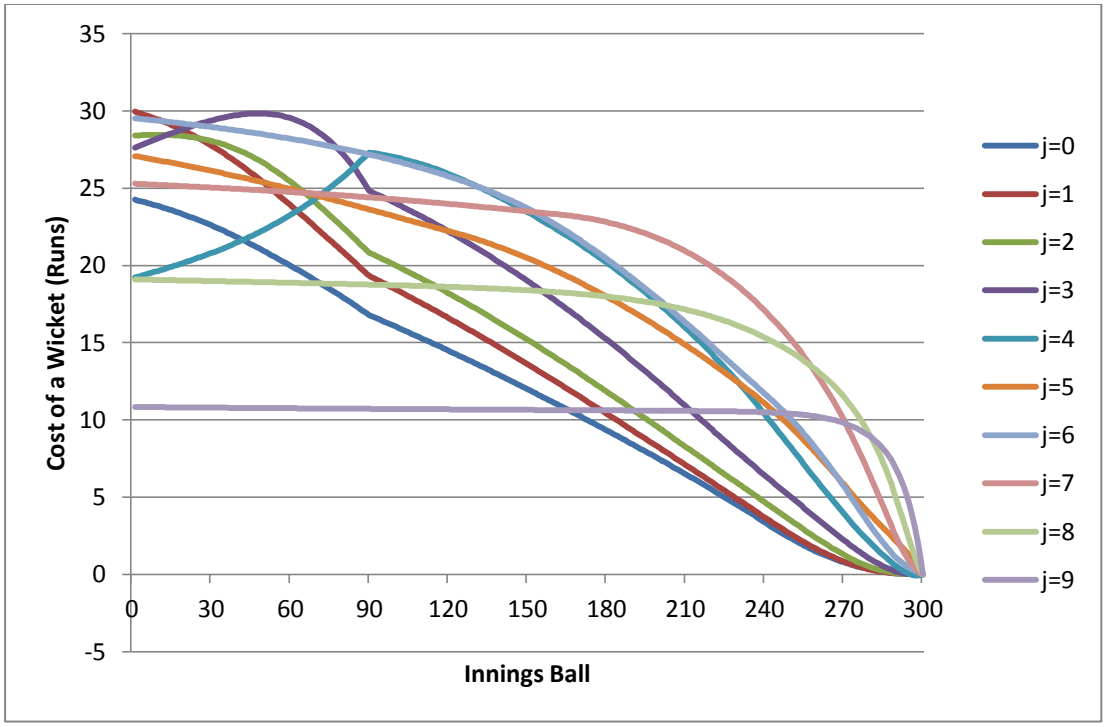
6.3.1 The cost-of-a-wicket functions

Recall that the key determinant of first-innings optimal strategy is $C(i, j)$, the cost of losing a wicket on ball i given that j wickets have already been lost. Having identified our expected additional runs functions $V(i, j)$, we can now estimate the cost of losing a wicket as

$$C(i, j) = V(i+1, j) - V(i+1, j+1)$$

We plot the $C(i, j)$ function for each value of j in Figure 6.3. It is clear that the relationship between $C(i, j)$ and each factor *i* and *j* is a rather complicated one, but as a general rule we can say that the cost of a wicket tends to decrease for any given *j*, as *i* increases, and tends to increase for any given *i*, as *j* increases, with this tendency being stronger in the second half of the innings. Note that in the early stages of the innings, however, the costs are not necessarily monotonic in the number of wickets lost; this is an indication of the complexity of the decision that players have to make on the field in terms of their risk strategies. One implication of this is the counter-intuitive result that the loss of a wicket can lead to it being optimal to increase the amount of risk taken.

Figure 6.3: C-functions in unknown conditions model, non-power-play era



Naturally, the cost of a wicket varies with the conditions, but the relative importance of conditions to the cost, depends on the number of wickets lost and the number of balls bowled. This is illustrated in Figure 6.4, which shows the C-functions for zero and six wickets lost for poor, approximately average and good batting conditions. It is clear that conditions are very important for determining the cost of a wicket near the start of the innings with six wickets lost (an unlikely situation), but not important in general with zero wickets down and, in either case, towards the end of the innings.

In better conditions, it is likely that two factors are approximately cancelling each other out. As a higher overall score is expected, each partnership is expected to contribute a higher number of runs than in worse conditions. This is counteracted by the lower probability of a team being bowled out within their 50 overs in better conditions. Clearly, this second factor does not apply when a team has lost six wickets very early on and this explains the importance

of conditions in that situation. Figure 6.5 compares the functions for three and eight wickets down and we see a similar pattern. The cost of losing your eighth wicket early in the innings is very large when conditions are good and smaller in poor conditions as less is expected from the last two partnerships.

Figure 6.4: Selected C-functions in various conditions, non-power-play era

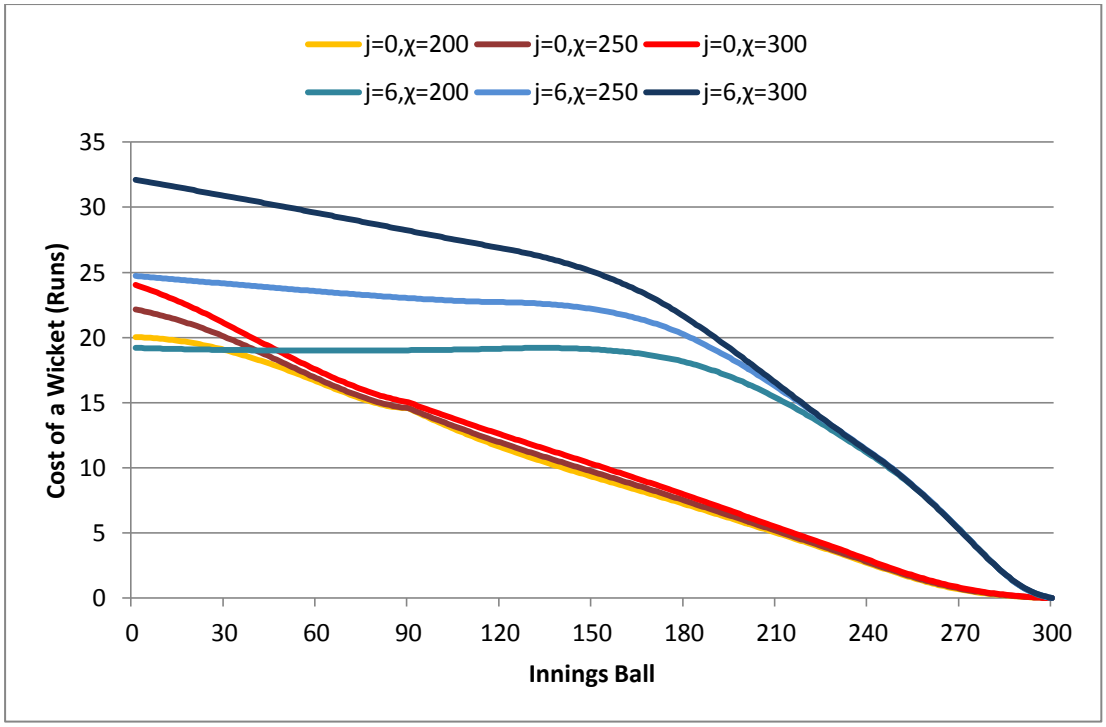
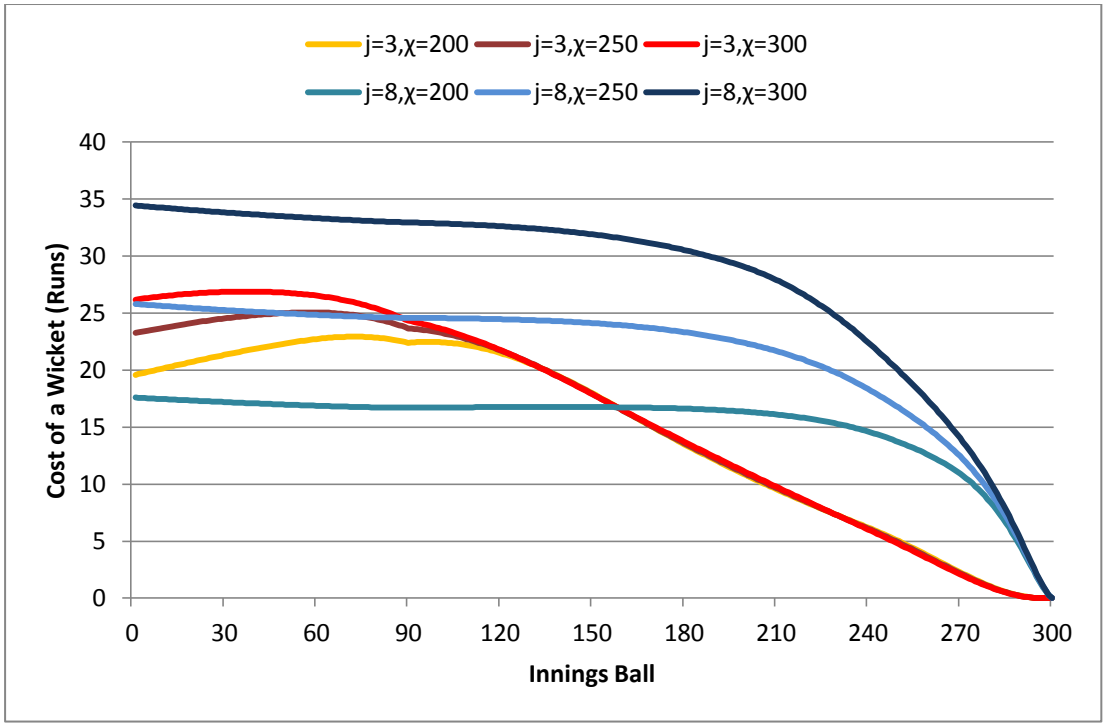


Figure 6.5: Selected C-functions in various conditions, non-power-play era



6.3.2 Inferring the Production Possibility Frontiers

Recall Equation (23)

$$\frac{\partial E\left[r_{ij}\right]}{\partial \eta_{ij}} = V^*(i+1, j+1) - V^*(i+1, j)$$

This equation implies that the expected additional runs function is maximised by choosing the level of risk where the absolute slope of the batsman's PPF is equivalent to the cost of a wicket. We propose that a rational batsman will adjust his chosen level of risk as the cost of getting out changes. Note that we are not assuming that a batsman will always choose to operate at the point on his PPF implied by the optimal level of risk; rather, we are suggesting that the cost of a

wicket provides a good reference point for us to compare a batsman's trade-off between scoring rates and the probability of survival. While a batsman is unlikely to be thinking about the exact equation of his PPF or the exact cost of a wicket at the time he is at the crease, we assume that he at least has a good enough idea about the amount of risk that is appropriate for the current game situation to choose a higher amount of risk for a lower cost of a wicket. That is, we assume that his risk adjustments to the changing cost of a wicket are monotonic in $C(i, j)$ and in the optimal direction.

Recall that we defined the runs function as $r(\kappa)$ and the probability of a wicket function as $\lambda(\kappa)$, noting at the time that we did not know the exact relationship between each variable and the risk parameter κ. Using the cost of a wicket as a proxy measure of the risk taken enables us to model these relationships and allows us to determine what a batsman is capable of by constructing his PPF. We achieve this by separately modelling expected runs, $E[r]$ and the probability of survival, η, as a function of the risk proxy, C, which yields an estimate for expected runs and the probability of survival for each possible value of C. The locus of points implied by the set of observed C values is our estimated PPF for that batsman.

Define $r(C)$ as a batsman's expected runs and $\eta(C)$ as his probability of survival, for a given cost of a wicket C. The inverse function of $\eta(C)$ is $\eta^{-1}(C) = C(\eta)$. The batsman's PPF is then defined by

$$r(\eta) = r(C(\eta))$$

6.3.3 The spline estimation procedure

We group our data by individual batsman and period of the innings (fielding restrictions or no restrictions). For each group we want to create two models linking first our runs variable r and second our survival variable η to our cost-of-a-wicket variable and conditions variable, C and χ respectively. For comparison purposes, not all the PPFs that we create contain this conditions variable. *A priori*, we expect that the cost-of-a-wicket variable will influence the runs variable negatively and the survival variable positively as batsmen should be relatively defensive when it is expensive to lose a wicket. We also expect that it should be easier to both score more quickly and survive in easier batting conditions.

The final thing needed to create the models is a functional form for the runs and survival functions. Before creating the models, we need to address the very important question of which functional form we should select for the model. Fitting the $r(C)$ and $\eta(C)$ curves through data means that we can use information from observed situations to get an estimate of a point on the PPF for any situation; however, we need to be careful with this. The PPFs give one the ability to determine the optimal point at which a batsman should operate, given the game situation, by setting the slope of the PPF equal to the slope of the indifference curve. The slope of the PPF function is very important and we are concerned that assuming a parametric functional form for two functions that are to be combined to form the PPF, may impose so much structure on the model that we may not get an accurate representation of the true PPF at any point. This is especially true as regression models are more concerned with the fit in the thick data areas than in the thin data areas, which is often justifiable on the grounds that the thin data areas are, by definition, rarely observed. In our case, the thin data area(s) of the PPFs are

thin because the batsman in question does not encounter the costs at which he chooses to operate in that region of his PPF often; however, it is perfectly possible that our PPFs might indicate that he should be operating in that thin data area much more often than he currently chooses.

SAS provides the capability to estimate Generalised Additive Models (GAMs). A GAM involves a new function of some form being added at each unique value of the independent variable, creating a smooth fit without a general functional form. In particular, we are able to construct semi-parametric models of the form

$$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + s(x_2)$$

The $s(x_2)$ represents a smoothing spline which is the solution to the optimisation problem

$$\min_{\eta} \sum_{i=1}^{n} (y_i - \eta(x_i))^2 + \iota \int_a^b (\eta''(t))^2 dt$$

where η is the set of all functions with two continuous derivatives, ι is a fixed constant indicating the size of the penalty for curvature in the function and *a* and *b* are the minimum and maximum values of x_i, respectively. The optimiser is a natural cubic spline with knots at the unique values of x - it is a piecewise cubic polynomial.

SAS contains a procedure PROC GAM that estimates the GAM. The resulting functions are flexible in that they do not have a specific functional form; however, we do need to determine the value of the smoothing parameter $\frac{\iota}{1+\iota}$. Setting ι very high implies a smoothing parameter of close to one, which means that curvature in the GAM is not tolerated and our

function approximates a straight line. Setting ι close to zero implies a very small smoothing parameter, which means that the GAM simply looks for the best fit to the data and the result is not a smooth function. Note that SAS does not allow the specification of ι directly; instead, it requires the specification of a degrees of freedom parameter, equal to the trace of the matrix of smoothing functions implied by a particular ι. We can specify any value greater than one for the degrees of freedom parameter.

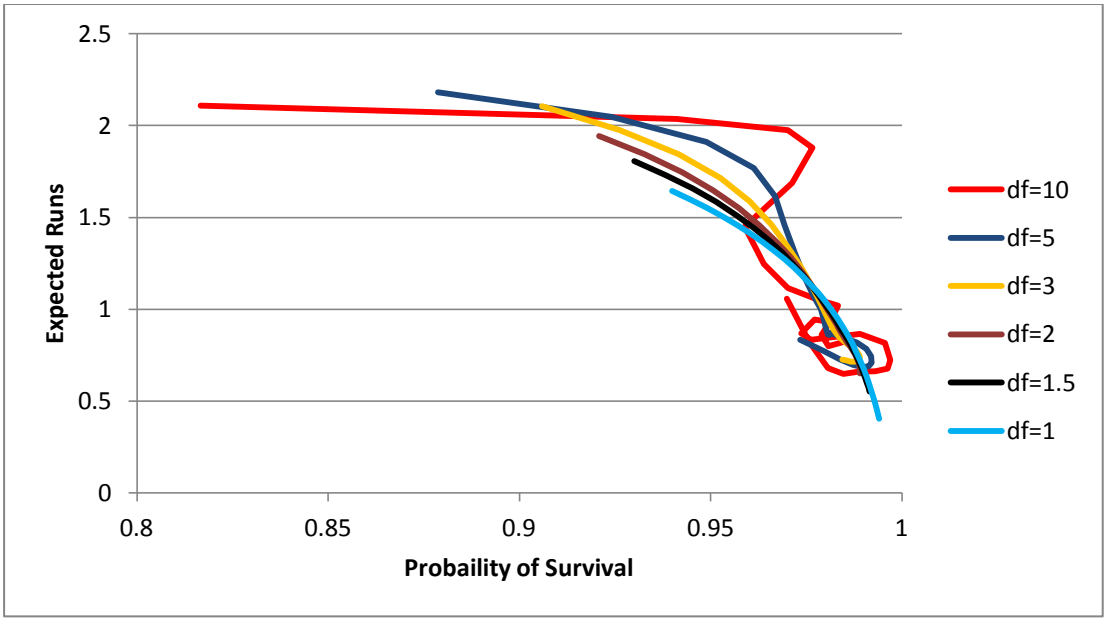
We create models for r and η where χ is a parametric term and C is a spline term. The most difficult aspect of the modeling process is choosing the best degrees of freedom parameter. If we choose a value too low, the resulting PPFs will fail to show the full curvature of the batsman's true underlying PPF. If we choose a degrees of freedom value that is too high, we will over fit our model and the PPF will contain curvature that does not really exist. SAS does contain an automatic selection criterion, the Generalised Cross Validation Function, which fits the model by leaving out one data point at a time and calculating the squared residual for that data point and summing these squared residuals, seeking the model that minimises this sum. This feature tended to over-fit our model in too many cases to be useful.

In order to investigate the most appropriate choice of degrees of freedom parameter, we try a number of values for three example batsmen in our non-fielding-restriction data set. We choose our batsmen for this analysis based on the frequency that they appear in the data, paying attention to both total number of balls faced and number of times dismissed. The latter is important as a wicket is a rather rare event; therefore, to construct a PPF we need to observe a batsman being dismissed in the data set a reasonable number of times. Andrew Symonds has faced the most balls of anyone in our data set (1567) and has been out 34 times. If our choice of

model cannot fit his data well, it would likely have little chance for many other players. Justin Kemp has faced 355 balls for ten dismissals, which is very thin data and therefore we want to investigate the performance of the GAM here too. Our third batsman is in the approximate middle of these two extremes - Brendon McCullum has faced 735 balls for 23 dismissals. Now that we are starting to look at individual batsmen, at this point it is important to note that our data contain, in most cases, a small snapshot of a player's career. Any conclusions about the strengths and weaknesses of individual players should be interpreted as being the case in our particular data set, but not necessarily extrapolated to form a judgement of that player in general.

In order to decide how many degrees of freedom we need for a reasonable fit, we construct the PPFs for our three example players where degrees of freedom is equal to one, 1.5, two, three, five and ten. Andrew Symonds' PPFs are shown in Figure 6.6, Brendon McCullum's in Figure 6.7 and Justin Kemp's in Figure 6.8. Note that for simplicity, these models have been constructed without the conditions variable.

Figure 6.6: PPFs for selected values of *df* – Andrew Symonds



Andrew Symonds' PPF is constructed from the most data. It is immediately clear that *df=10* and *df=5* suffer from substantial over fitting, while *df=3* looks better but suffers from an unexpected shape near the high-survival end point. There may be cricket reasons to explain some unexpected results, so we need to be careful to not use such results as a reason to over-smooth the data. Overall, *df=2* has the largest *df* value for which the shape looks reasonable. For Brendon McCullum, we remove *df=10* and *df=5* in order to improve the clarity of the graph. We see here that *df=3* has a strange feature in the high-survival region, while *df=2* has a slight concavity of the production set in the same region and *df=1.5* has no such features.

Figure 6.7: PPFs for selected values of *df* – Brendon McCullum

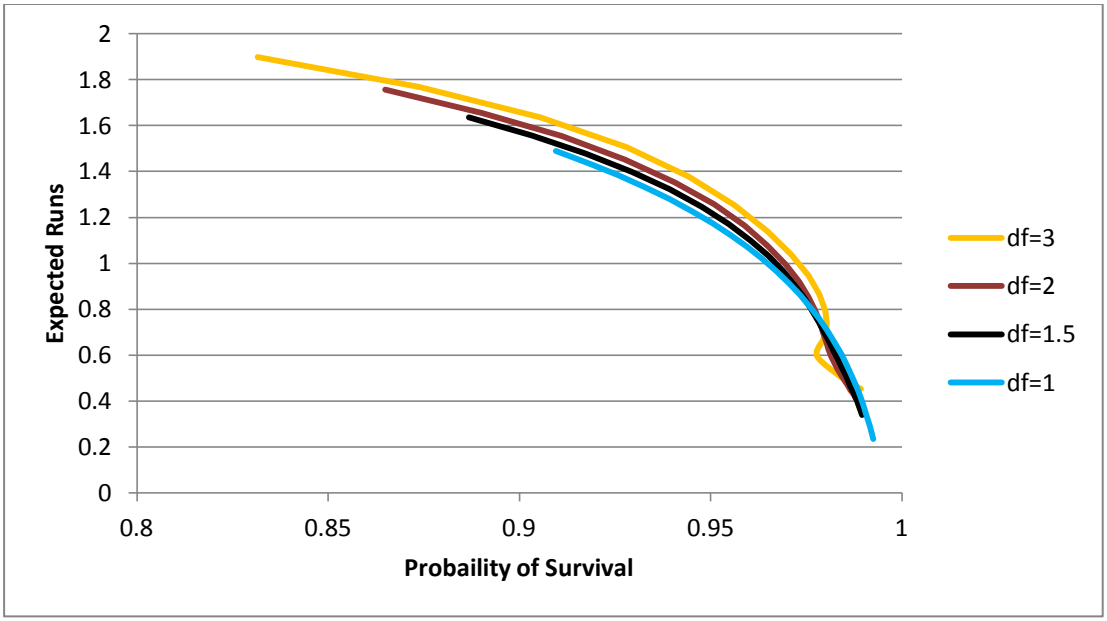
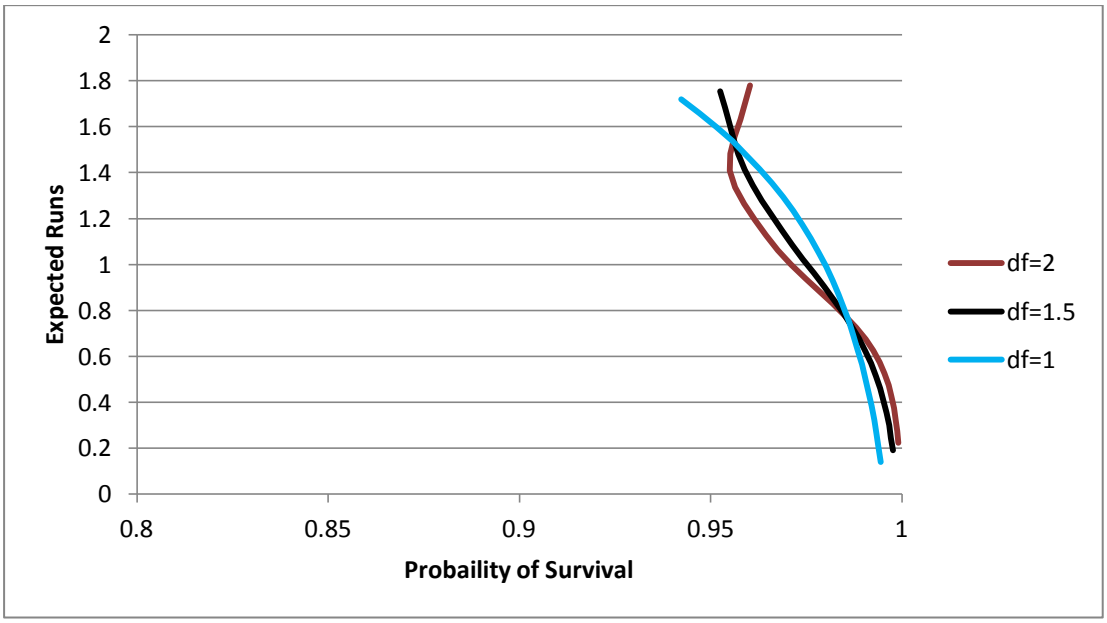


Figure 6.8: PPFs for selected values of *df* – Justin Kemp



Justin Kemp's PPFs, which are based on the least amount of data of the three players, show an unusual feature at the high-scoring end for *df=2*, where the graph implies that he can

increase his scoring rate to its highest point and increase his probability of survival as he does so. This seems highly unlikely. The graph for *df=1.5* looks reasonable, although we have a concavity-of-the-production-set region.

Taking the above findings into account and erring on the side of our prior belief that the PPFs should have a weakly convex production set and exhibit monotonicity in the survival variable, we decide on the following rules for the selection of the degrees of freedom parameter.

$$df = \begin{cases} 1.2, & 351 \le n \le 500 \\ 1.3, & 501 \le n \le 1000 \\ 1.5, & 1001 \le n \\ N/A, & \text{Otherwise} \end{cases}$$

No PPF is estimated for players with 350 or fewer observations. We note that the goal of this overall modeling strategy is that we want it to be possible for our models to show concavity of the production set (both local and global) or non-monotonicity, but we also want to make it difficult for these factors to appear in the model by chance alone, meaning that the model must have a high degree of confidence that these factors exist before they appear this way in the results.

6.3.4 Estimating the PPFs

Having obtained, from our GAM models, all the necessary information from which to construct our PPFs, we outline the relevant functions of one of New Zealand's longest-serving batsmen, Chris Harris. Figures 6.9 and 6.10 contain the expected runs and probability of

survival functions, respectively, with the cost of a wicket as the dependent variable. For simplicity, these are models constructed from the unknown conditions data set. These figures show that this batsman scores relatively slowly and has a higher probability of survival when the cost of a wicket is high, which is a rational strategy. Harris' expected-runs function is convex while his survival function is concave. It is the degree of this convexity and concavity that determines whether the overall production set will be convex or not. These curves are estimated only over the range of costs observed for Chris Harris in the data set. We are interested in his batting ability as revealed by the data; therefore, we do not make any assumption about what he could achieve over extended ranges of cost of a wicket.

Figure 6.9: Expected Runs function for Chris Harris, Unknown Conditions

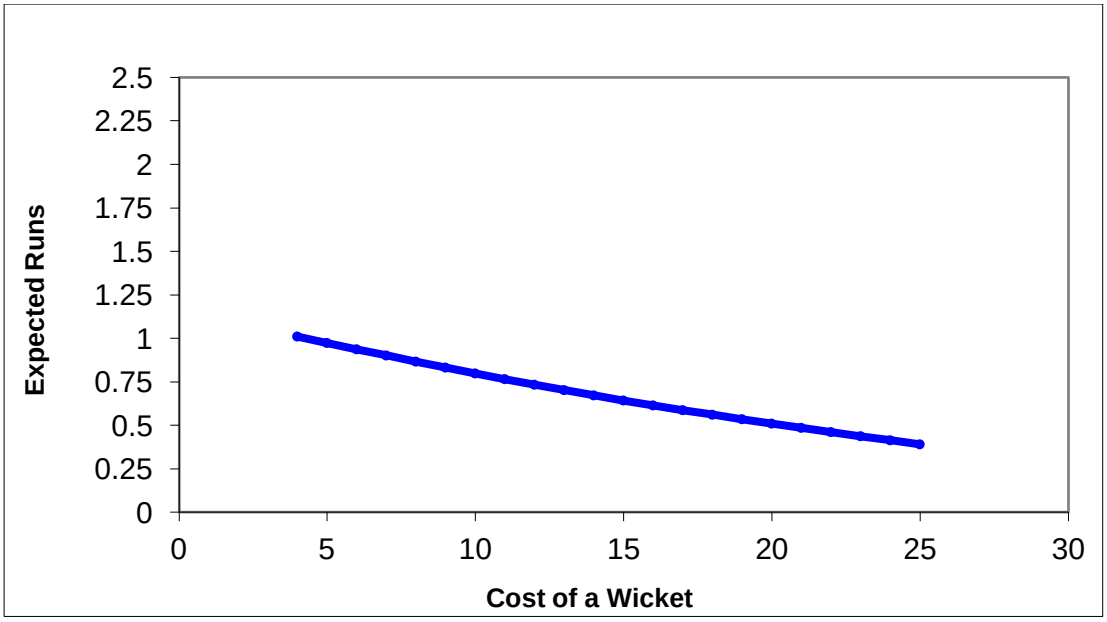
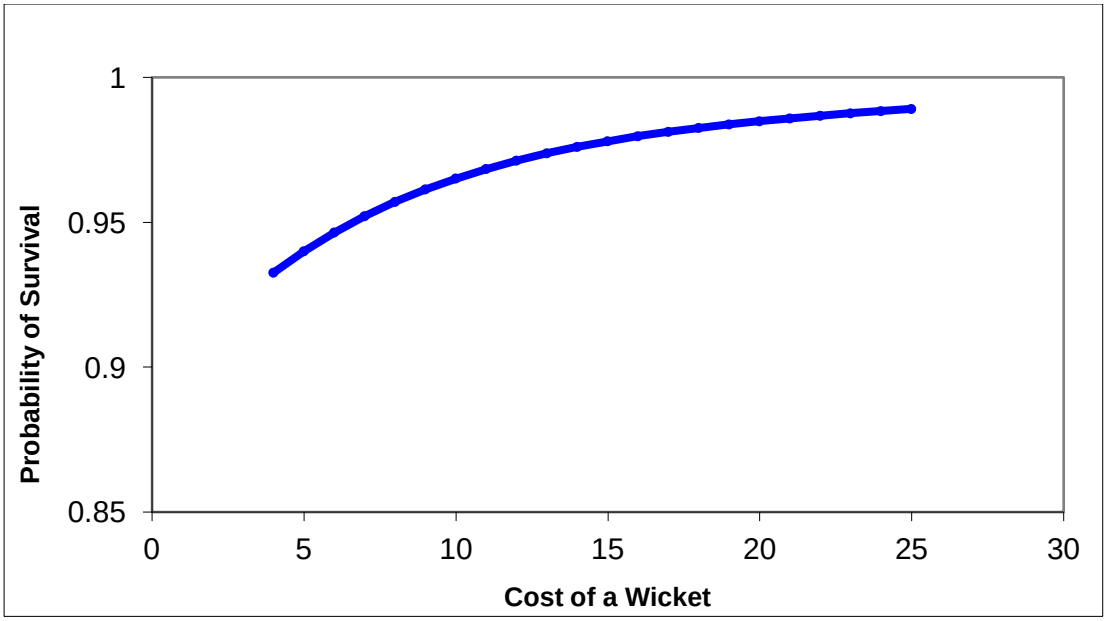
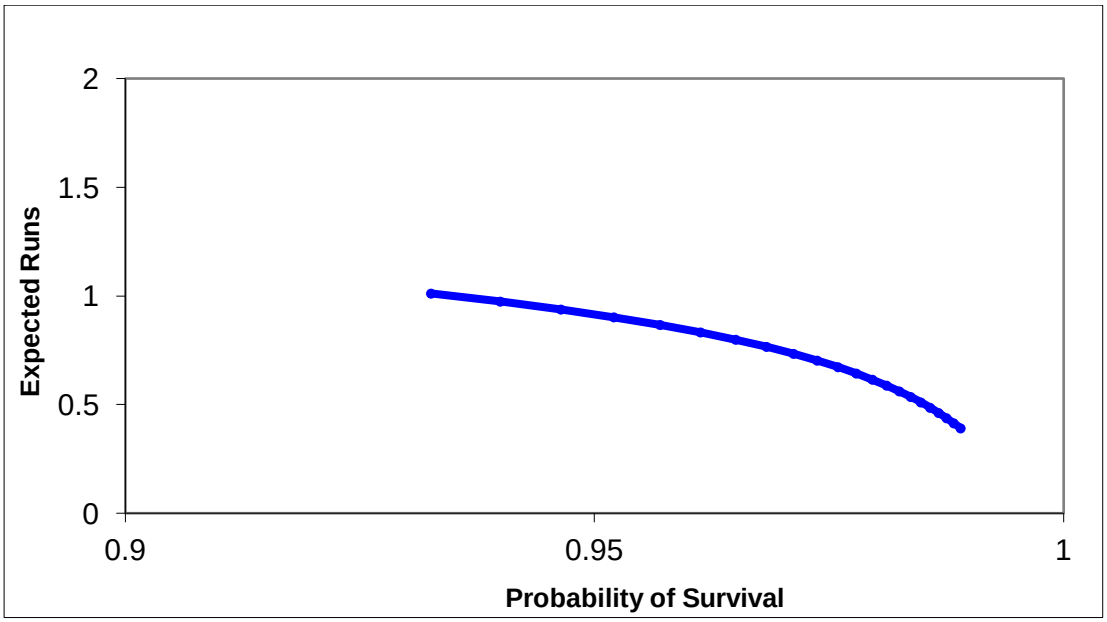


Figure 6.10: Survival function for Chris Harris, Unknown Conditions



Putting the data in Figures 6.11 and 6.12 together, for each cost of a wicket, yields the PPF for Chris Harris. This PPF reveals the locus of points from which Harris can select his strategy.

Figure 6.11: PPF for Chris Harris, Unknown Conditions

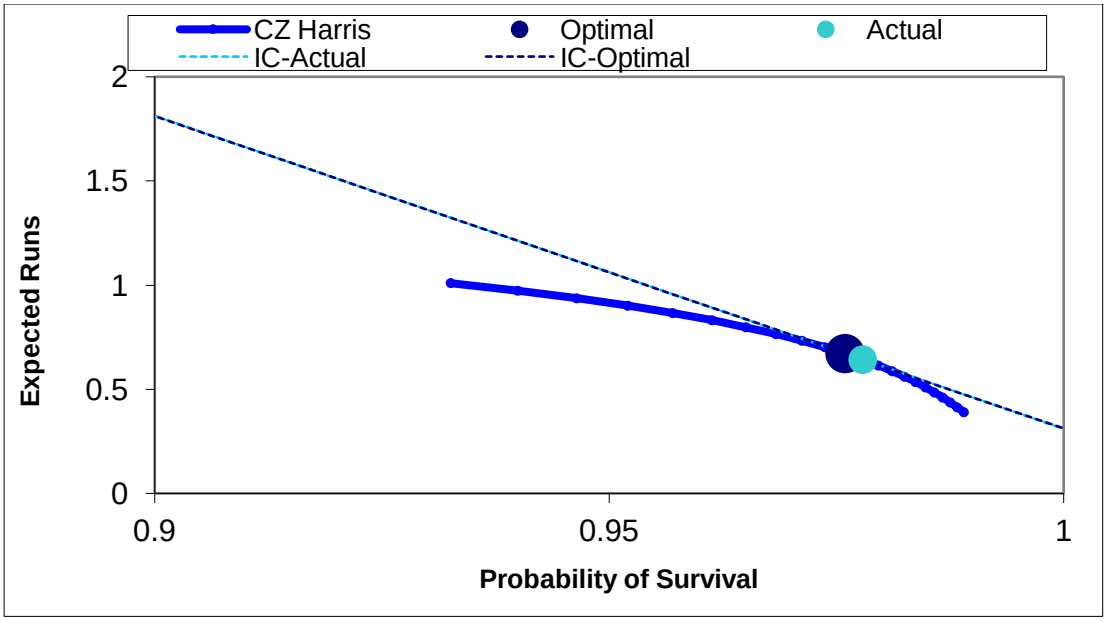


Once we have established the PPF of a batsman, we are able to determine their optimal strategy, given that PPF, in any game situation. Recall the solution to our optimisation problem, given in Equation (23)

$$\frac{\partial E[r_{ij}]}{\partial \eta_{ij}} = -C(i, j) \tag{23}$$

A batsman should operate where the slope of his production function is equal to the negative cost of a wicket. Since the cost of a wicket is a constant, given a particular game situation (i, j), a batsman has linear indifference curves with slope equal to $-C(i, j)$ with increasing preferences in both expected runs and survival probability. Our PPFs were constructed by matching expected runs and survival probabilities for given values of $C(i, j)$; therefore, we can infer the strategy chosen by a batsman for a particular $C(i, j)$ and compare it to the optimal strategy given his PPF. We show the case of Chris Harris, in unknown conditions, where the cost of a wicket is equal to 15, in Figure 6.13. The larger, dark blue dot represents the optimal point for Harris, given this cost, while the smaller, light blue dot represents his chosen point. In this situation, the chosen point is very close to the optimal point and Harris is approximating the optimal strategy. The light blue indifference curve, which his actual choice places him on, is almost indistinguishable from the dark blue indifference curve, on which his optimal choice would place him.

Figure 6.13: Optimisation for Chris Harris, Unknown Conditions, $C = 15$



It would be useful to be able to determine the impact of a suboptimal choice in terms of value function for expected additional runs. Recall our value function

$$V(i, j) = E[r_{ij}] + (1-\eta_{ij})V(i+1, j+1) + \eta_{ij}V(i+1, j) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

We define the value function for a particular individual batsman, B, in state (i, j) and taking risk level κ. We write

$$V(i, j \mid B_\kappa) = E[r_{B_\kappa}] + (1-\eta_{B_\kappa})V(i+1, j+1) + \eta_{B_\kappa}V(i+1, j) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}}$$

$$\Rightarrow \quad V(i, j \mid B_\kappa) = E[r_{B_\kappa}] + \eta_{B_\kappa}C(i, j) + V(i+1, j+1) + \frac{\gamma_{ij}\tau_{ij}}{1-\gamma_{ij}} \qquad (24)$$

The third and fourth terms of the RHS of Equation (24) are constants in state (i, j), as they relate to the performance of the average team. The first and second terms contain important

information about the performance of batsman *B*. Define the Expected Gross Contribution (EGC) of batsman *B*, taking risk level κ, as

$$EGC(B_\kappa) = E[r_{B_\kappa}] + \eta_{B_\kappa} C(i, j)$$

The EGC measure enables us to make three important comparisons. First, it enables us to compare the chosen strategy κ with the optimal strategy κ^*, for a given batsman. Second, it enables us to compare the performance of two batsmen *B* and *D*, assuming either that each batsman operates under their chosen strategy B_κ and D_κ, or their optimal strategy B_{κ^*} and D_{κ^*}. Third, it enables us to compare the outcome from a batsman's chosen or optimal strategy with the average outcome as estimated in the dynamic programme for the V-functions. This last comparison is outside the scope of this thesis.

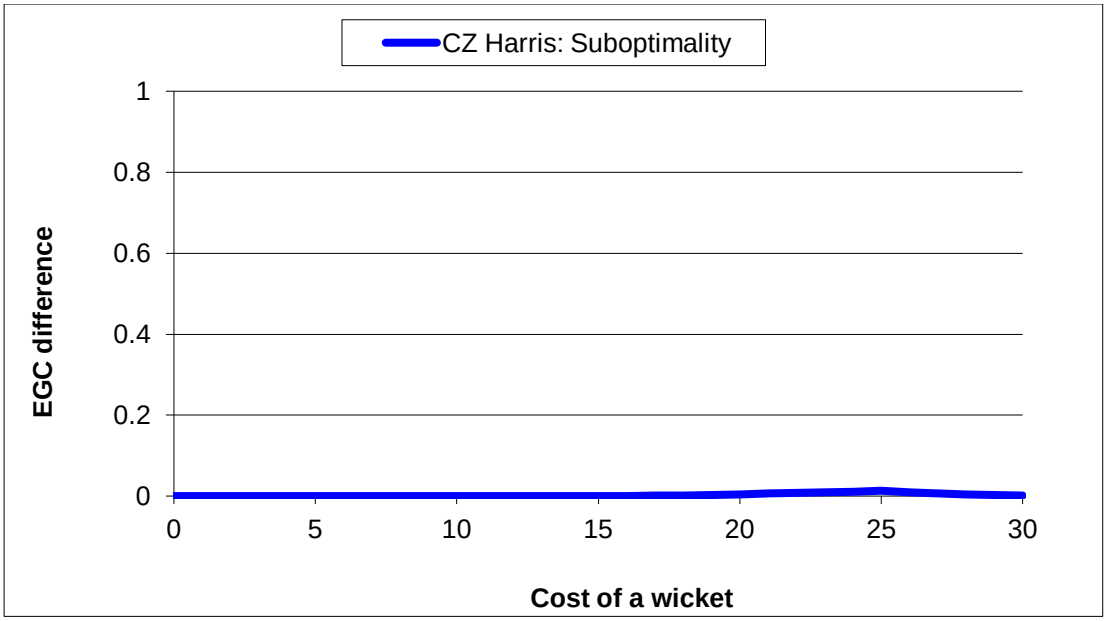
6.3.5 Illustrative examples of the use of the PPFs and EGC measure

This section contains several examples illustrating the information that one can extract from the PPFs and the EGC measure. Returning to our example of Chris Harris in Figure 6.13, Harris' chosen strategy where the cost of a wicket is equal to 15 yields $E[r_{Harris_\kappa}] = 0.643$ and $\eta_{Harris_\kappa} = 0.978$. If he followed the optimal strategy he would have operated at the point where $E[r_{Harris_{\kappa^*}}] = 0.672$ and $\eta_{Harris_{\kappa^*}} = 0.976$, a slightly more aggressive point. Given the cost of 15 runs, $EGC_{Harris_\kappa} = 15.31205$ while $EGC_{Harris_{\kappa^*}} = 15.31206$. These values are almost identical, indicating that Harris' small amount of sub-optimality is doing very little harm in this situation. The EGC measure is particularly useful because in some situations a batsman may be operating

far away from his optimal point in terms of the Euclidian distance on his PPF, but his poor choice may have a very small impact on the resultant V-function. This is generally the case where a batsman has a relatively flat PPF in the relevant range of risk choices and the cost of a wicket is high. In other situations, operating slightly away from the optimal point may have a very large impact on the resultant V-function. This generally occurs when a batsman has a steep PPF in the relevant range and the cost of a wicket is low.

To this point we have compared Chris Harris' chosen strategy to his optimal strategy for a single value cost of a wicket, $C = 15$. It would be useful to illustrate the difference between these strategies for all possible costs. It is possible to do this using the EGC measure, by calculating the difference between the EGCs implied by the optimal strategy and the chosen strategy for each cost of a wicket. The difference between two EGC functions can be interpreted as the difference in runs of the V-functions implied by the two strategies. Note that this assumes that each strategy is chosen for the current ball only; the subsequent V-functions are taken as given. We show the difference between the EGC functions for the optimal and actual strategies, again using the example of Harris, in Figure 6.14. It is clear that this batsman is achieving very close to the optimal outcome, for all costs. We will show in the subsequent analysis that many other batsmen are not so strategically aware.

Figure 6.14: Optimal vs. chosen strategy for Chris Harris, Unknown Conditions



To this point we have shown the unknown conditions PPFs. By including our conditions variable, χ , as an explanatory variable in our GAM models, we are able to estimate PPFs for different batting conditions. Note that the conditions variable enters the model parametrically; therefore, our resultant model is of a semi-parametric nature. Including a conditions variable has two benefits. First, it enables the more accurate modeling of the relationship between the cost of a wicket and the scoring rate and survival rate due to the removal from consideration of the variation in each dependent variable attributable to conditions. Second, we are able to see the impact of conditions on the PPFs of individual players and investigate the different strategies that they should play under different conditions. Figure 6.15 shows the PPFs for Kumar Sangakkara under very poor and very good batting conditions. It is clear that in the better conditions he can score at the same rate with a much greater probability of survival. Since this is true for all values of expected runs in the range of

the PPFs we can say that Sangakkara is unambiguously more capable in good batting conditions than in poor batting conditions. This statement, however, assumes that he adopts the optimal strategy. We show the difference in the EGC functions implied by the optimal and actual risk choice for each value of conditions in Figure 6.16. It is clear that some of Sangakkara's higher capability under good batting conditions is lost due to a higher level of sub-optimality. This might indicate, for example, that Sangakkara does not take enough risks when conditions are good for batting.

Figure 6.15: PPF for Kumar Sangakkara under different conditions

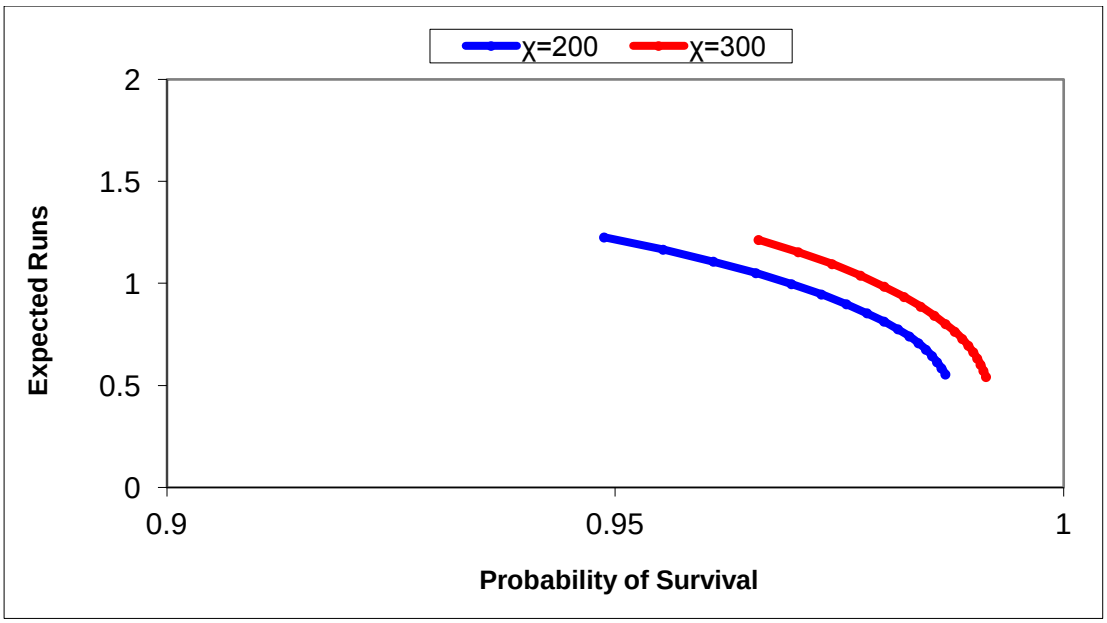
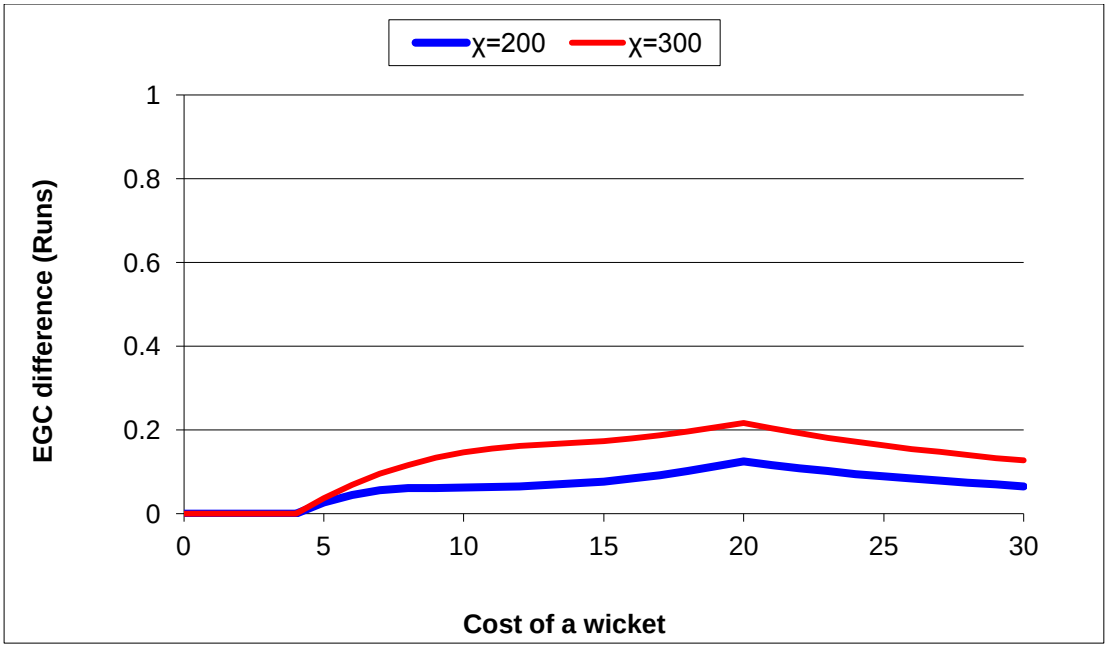


Figure 6.16: Optimal vs. Chosen strategy Sangakkara under different conditions



In addition to determining the PPF of a batsman under different conditions, it is also possible to include a variable for the number of balls faced by a batsman in their current innings. This assists with the identification of slow starters and can suggest different optimal strategies depending on how long the batsman has been present at the crease. We do, however, have a lower degree of confidence about these functions as it is perhaps too many variables to include when dealing with a relatively small data set for each batsman. We therefore exclude this addition from the analysis in this thesis but note that, with sufficient data, this is a simple addition to the modeling process.

Our analysis so far has looked at the abilities of individual batsmen as well as the impact of their choices. One of the most useful applications of these PPFs is that they enable many comparisons between any two individual players. We can investigate whether one player has unambiguously greater ability than another, or whether this is situation-dependent. In

addition, we can compare degrees of optimality in order to answer the question as to whether a player with greater ability is always a better choice.

The first player comparison is between two middle-order batsmen, Michael Hussey and Mark Boucher. Their PPFs are shown in Figure 6.17. We see that Hussey's PPF dominates Boucher's in every situation, which implies that Hussey has greater ability. In Figure 6.18 we show the difference in the EGC functions for each player under the optimal strategy and under the actual strategy. By convention, these are shown as the EGC of the blue batsman from the PPF graph minus the EGC of the red batsman; therefore, any positive difference implies an advantage to Hussey over Boucher in this case. It is clear from Figure 6.18 that, while Hussey has a substantial advantage over Boucher in terms of ability, he suffers from much greater sub-optimality and the difference between the players in terms of what they actually achieve is lower than it might otherwise be if both players behaved optimally.

Figure 6.17: PPFs for Hussey and Boucher, $\chi = 250$

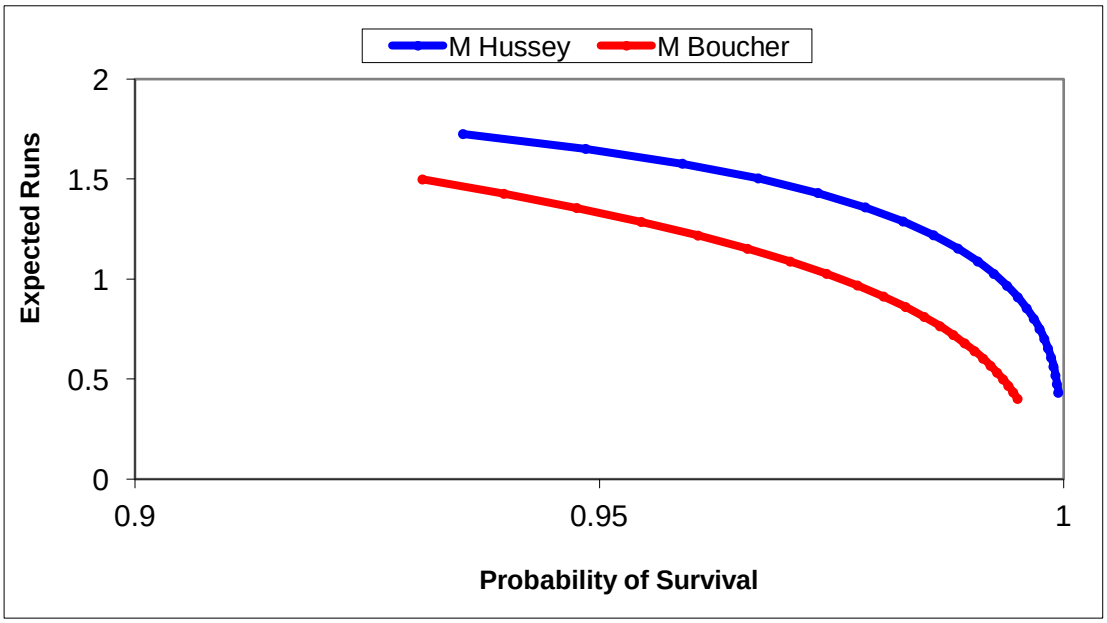
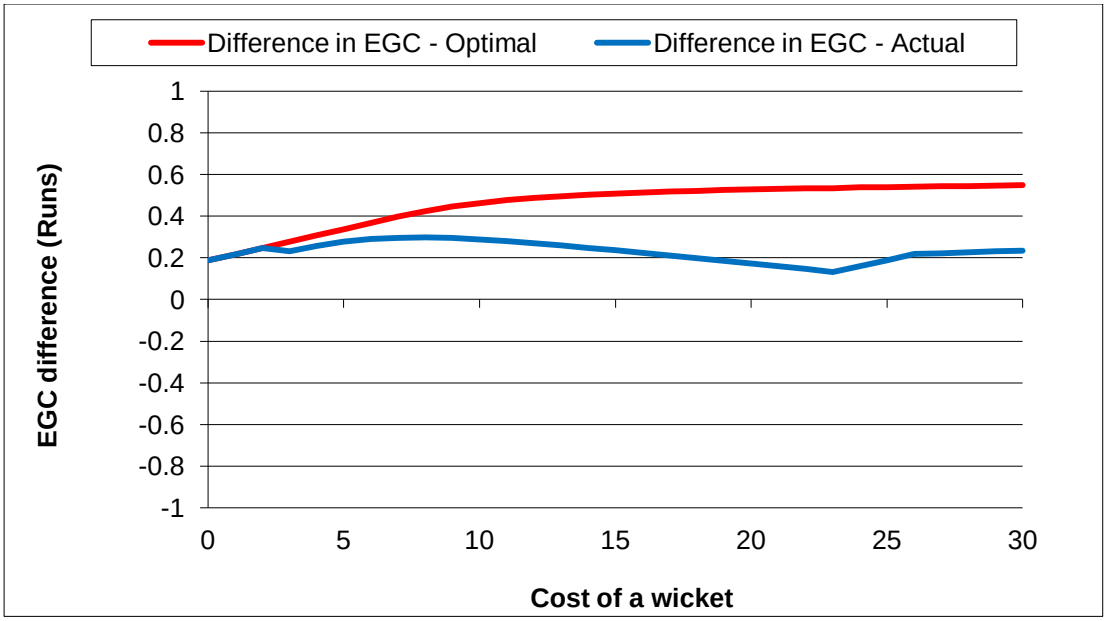


Figure 6.18: Difference in EGC functions for Hussey and Boucher, $\chi = 250$



Our functions show that it is often not the case that one batsman is better than another in all situations. Consider the comparison of Yuvraj Singh and Herschelle Gibbs, whose PPFs are shown in Figure 6.19. These functions have an intersection point, meaning that for some survival probabilities, Singh is able to score more quickly than Gibbs while for others the situation is reversed. Figures 6.20 and 6.21 show the chosen and optimal strategies when the cost of a wicket is 5 and 20, respectively. In Figure 6.20, despite Gibbs operating exactly at his optimal point, Singh operates on a higher indifference curve due to his greater ability. For the higher cost of 20 runs in Figure 6.21, Gibbs is the better choice.[14]

[14] Note that the decision of who is the better choice is complicated somewhat by the fact that a high-cost situation tends to turn into a low-cost situation if wickets do not fall.

Figure 6.19: PPFs for Singh and Gibbs, $\chi = 250$

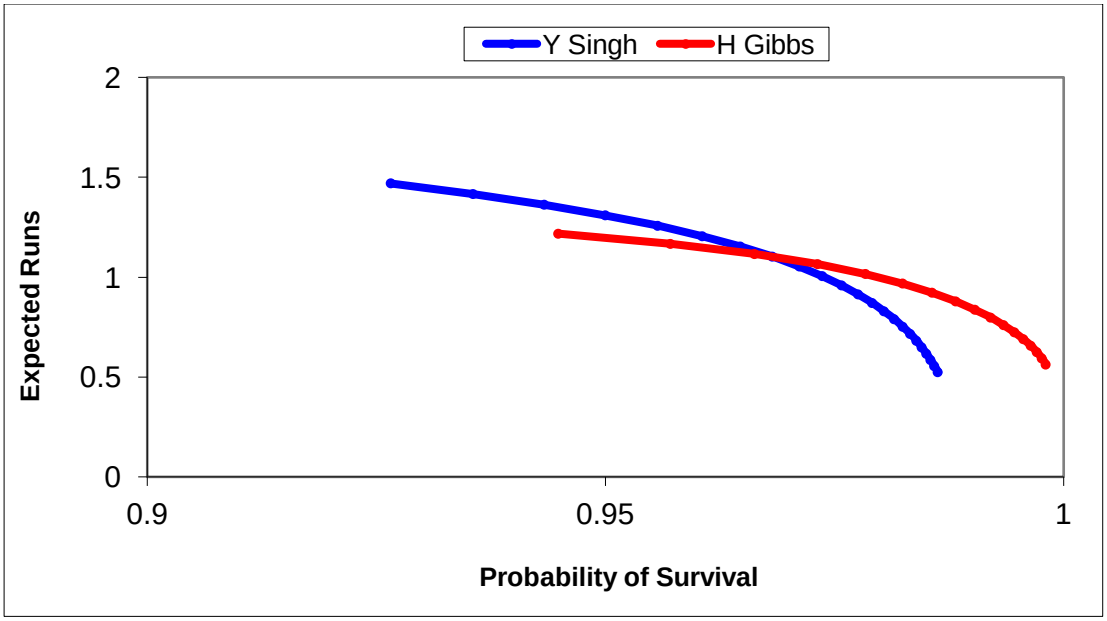


Figure 6.20: Optimal and chosen strategies for Singh and Gibbs, $\chi = 250$, C=5

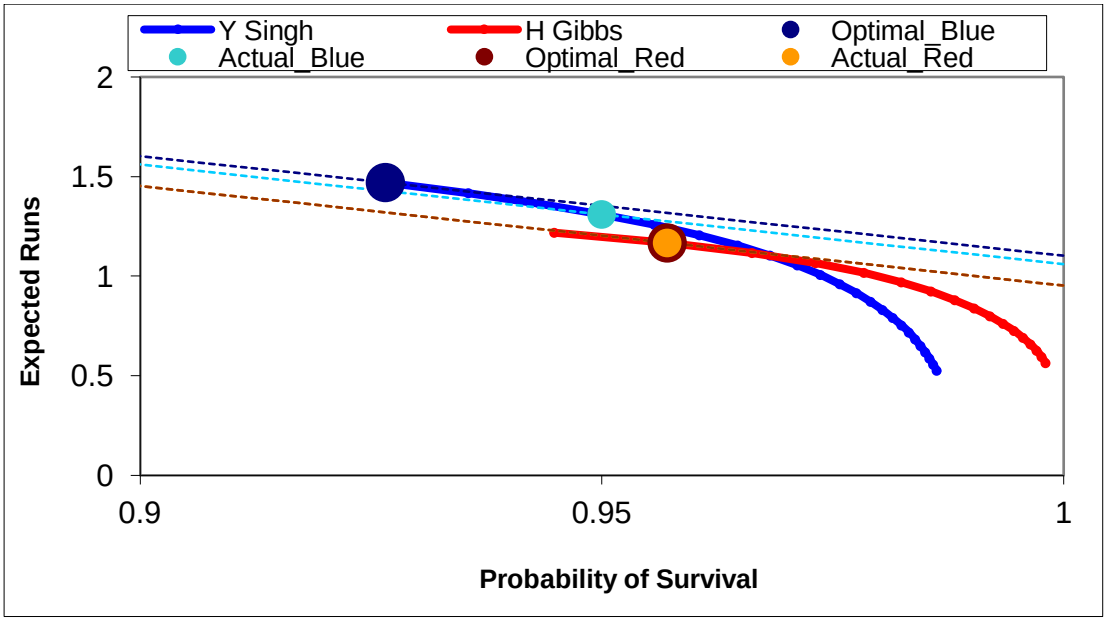
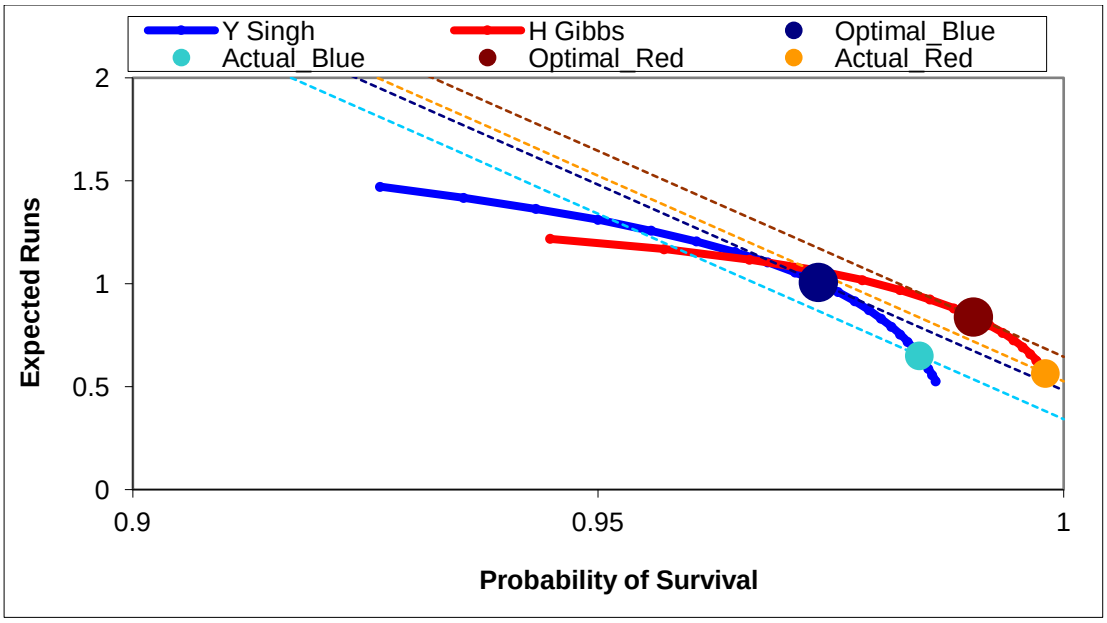
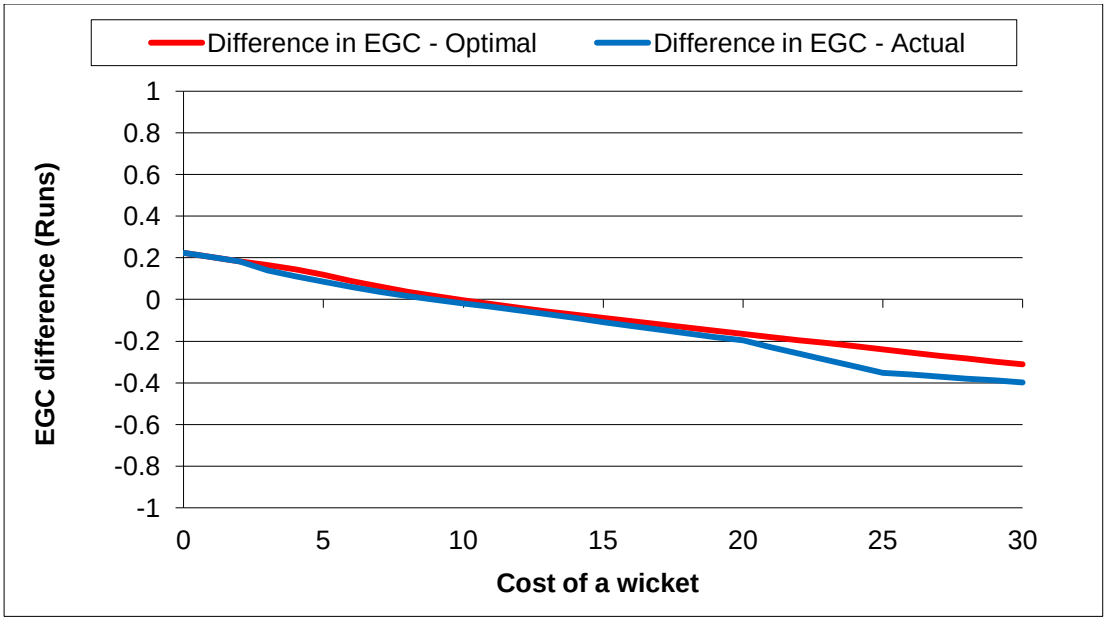


Figure 6.21: Optimal and chosen strategies for Singh and Gibbs, $\chi = 250$, C=20



We show the comparison between Singh and Gibbs more generally in Figure 6.22. The red line indicates that, assuming both batsmen play optimally, Singh is the better choice for situations where the cost of a wicket is less than approximately ten. For costs of greater than ten Gibbs takes over as the better choice, which is implied by the red line going below zero. For most costs there is not a great deal of difference between the red "optimal strategy" line and the blue "actual strategy" line. This does not mean that the batsmen are being optimal; rather, it implies that they are operating with similar levels of sub-optimality. For very high costs Singh starts to be substantially more sub-optimal than Gibbs.

Figure 6.22: Difference in EGC functions for Singh and Gibbs, $\chi = 250$



6.3.6 The fielding restrictions period and the PPFs

In our analysis we built separate PPF functions for the restrictions and non-restrictions periods of the first innings. The examples given have exclusively come from the non-restrictions data. Unfortunately our estimation process did not result in well-behaved PPFs for the restrictions period, except in a few cases. The vast majority of non-restrictions period PPFs exhibited the expected features of monotonicity in survival and the convexity of the production set, although there were a few exceptions.

Where a function could not be properly estimated, there are a few possible cricket explanations as to why the cost of a wicket was not a good indicator of the risk that the batsman was taking. First, a batsman may not be aware that he should take the cost of a wicket into account when determining his risk strategy. Given that the procedure works extremely well in

the “without restrictions” period, this is an unlikely explanation. Second, a batsman may be aware that he should consider the cost of a wicket, but he may not have a good idea about what the cost of a wicket is in any given game situation. This is particularly likely to be the case in the restrictions period, where the costs do not vary to the great degree that they do in the no restrictions period. Third, a player may be well aware of both the need to consider the cost of a wicket and its approximate current value, yet chooses to ignore this information. This is a possibility, as the potential scoring methods are reduced for batsmen when the field is close, as they often need to hit the ball over the top of the fielders, as finding a gap along the ground is more difficult. Some batsman may prefer to play purely on instinct or by the old cricket adage of playing every ball on its merits.

We have developed an alternative method for constructing the PPFs in the restrictions period, involving assigning a level of risk to particular cricket shots played. This is useful in that it gives us a complete set of functions in order to run full innings simulations; however, this is outside the scope of this thesis.

6.4 Concluding remarks

The PPFs can be applied in a range of ways, from determining players’ abilities and assessing how well they use it, to determining which players are better in which conditions and in which game situations, to potentially answering questions about what a particular batting line-up following a particular strategy might be able to achieve, in both average and optimal worlds.

CHAPTER 7

Determining the winner of an abandoned match

7.1 Introduction

In an ideal world, players and fans alike would be able to enjoy a full-length contest of 300 balls per innings, on every occasion that a match is scheduled. Unfortunately, this is not always possible due to wet weather. The mechanics of bowling a cricket ball, which usually involve a momentum-building sprint, followed by a jump, and culminating in the arm rotating at pace as the ball is delivered, mean that it is considered too dangerous to play when the ground is slippery. As a result, play does not occur during a period of significant rain and, depending on the severity and duration of the rain, for a period after the rain stops as the ground must first be deemed dry enough for play. It is also possible for bad light to prevent play; however, this has become a rare event in the modern game as most stadiums are equipped with floodlights, often for the specific purpose of enabling games to be played at least partially at night.

There are several possible approaches for cricket authorities to take in determining what to do when part of the playing time is lost in a match. One option is to complete the match the following day; however, this does not fit well with playing schedules (which are often very busy), television coverage and spectator interests. A second option is to abandon the match completely, declaring neither team as the winner (in cricket terminology, this is called a no-

result). Consider the situation where a match has almost been completed and one team is almost certain to win. Declaring a no-result would be very unfair. A third option is to have some rules in place that enable a shorter match to be played. This is the approach that is almost exclusively used in ODI cricket today and this chapter focuses on the fairness of various sets of rain rules.

There are three main game situations to consider when thinking about the impact of a weather interruption. First, bad weather may delay the start of the match. Second, there may be an interruption which occurs after the start of the match and is short enough to enable further play after the interruption. Third, an interruption may occur which is so severe or so close to the end of the match that no further play is possible after the interruption. We note that it is possible for any combination of these interruptions to occur in any given match, as multiple interruptions do happen.

The situation where rain delays the start of the match is relatively trivial. In this situation it is very easy to come up with a solution that is fair to both teams. Since the amount of time lost is known at the start of the match, each team's batting innings can be reduced by the same number of overs. For example, if half the playing time were lost, then each team would bat for 25 overs, rather than 50.

If bad weather causes an interruption in play but there is enough time for more play after the interruption, we face a situation where it is far more difficult to invoke a rule which makes it fair for both teams. Most obviously, if the interruption occurs during the second innings of the match, Team 1 will have already batted for 50 overs; therefore, all the lost overs must be taken away from Team 2's innings. First-innings interruptions are also unfair. Consider the situation where Team 1 bats for 30 overs, then it rains and by the time the match is able to

be resumed there is time for just 30 further overs. Team 1 is disadvantaged if Team 2 are asked to bat for 30 overs and attempt to beat the score that Team 1 achieved in their 30 overs. This is because Team 1 would have been assessing their optimal risk strategy under the assumption that they had 50 overs to spread their ten wickets over, resulting in more defensive play than if they knew *ex ante*, as Team 2 would know, that they would only have 30 overs to bat. It is necessary for a fair rule to adjust the target score of Team 2 in these situations.

Sometimes bad weather can bring about the end of the match. This might be because the rain lasts for many hours, is so heavy that it takes hours to get the ground dry again after the rain stops, or we are close to the end of the match when the interruption occurs. In this situation, we need a fair rule to decide the winner of the match. We note that this is simply a special case of the previous situation, as each rule that we investigate sets a target score for Team 2 even in the event of no more play. If Team 2 are in excess of this target score at the time of the abandonment of the game then they are declared the winner.

Throughout this chapter we discuss several of the target-adjustment rules (rain rules) that have been used over the history of ODI cricket and we propose our own alternative rule. In order to compare the different rules, we introduce a method with which we can assess the accuracy of each one. Using this approach we show that our proposed rule is fairer than the Duckworth/Lewis (D/L) rule currently sanctioned for use in ODIs. Furthermore, we comment on the key differences between our rule and the D/L rule and assess the importance of each difference.

7.2 The rule-assessment procedure

A good rain rule should, as accurately as possible, answer the question "in the current situation, which team is more likely to win this match?" It is important to note that it is impossible to use matches that were actually interrupted or abandoned due to bad weather to assess the fairness of a rain rule. By definition, we do not know who the truly deserving winner of these matches was, as they could have won or lost partly due to an unfair rain rule. Therefore, we take the set of completed, uninterrupted matches and create artificial interruptions in these matches and we assume that each interruption results in the abandonment of the match.

We use the same data set of 311 matches as we used in Chapters 5 and 6; however, there is one game in this data set where the second-innings information is incomplete. In addition, there are two ties in our data set. Since our analysis investigates the ability of various rain rules to predict the winner of a match, these matches add very little value and unnecessarily complicate matters; therefore, we exclude them from our data set. This leaves us with 308 remaining matches in which we create artificial interruptions.

At each artificial interruption, we determine the winner predicted by each rain rule by calculating the revised target according to the rule and looking at whether Team 2 is ahead of that target or not. Since these were, in fact, uninterrupted matches, we know which team went on to win each match; therefore, we can calculate the percentage of winners correctly predicted by each rain rule. We call this measure the Correct Prediction Percentage (CPP). The more interesting use of the rain rule occurs when the players are able to continue the match after an

interruption and a revised target is calculated; however, we cannot observe both the result with and without an interruption. The purpose of the CPP measure is to determine the most accurate method, which could then be used to calculate the revised target score in an interrupted match where a resumption in play is possible.

The rules of ODI cricket require that at least 20 overs must be faced by Team 2 to constitute a match. Matches that are abandoned before reaching this threshold are deemed to be no-results. As the minimum number of overs to constitute a match has varied over the history of ODI cricket, we illustrate the differences in CPP for all potential abandonments in Team 2's innings in order to provide a complete picture of each method's accuracy.

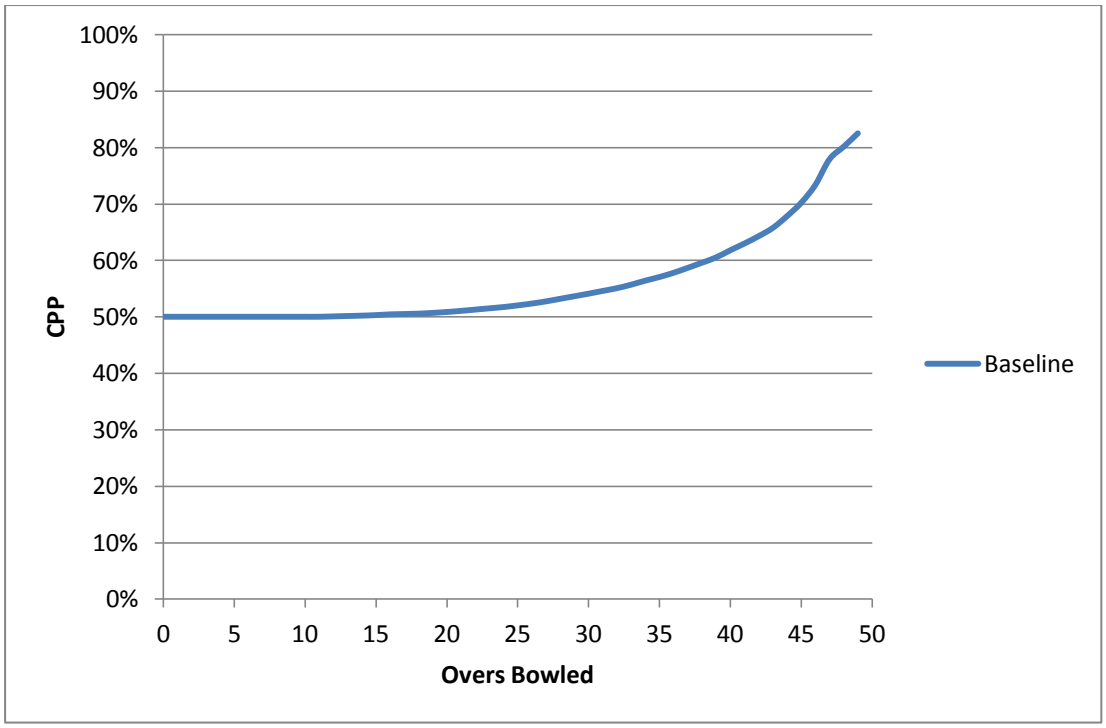
We create our artificial abandonments at the end of each over, from the first until the 49th and we plot the CPP for each method. In almost every match in which Team 2 is the winner and in some where Team 1 is the winner, Team 2's innings does not last the full 50 overs, due to their either achieving the target or being bowled out. In our assessment measure, we count any match that has already been decided as a correct prediction of any rain rule in order to show the likely impact of the choice of rule in any given match where, ex ante, we do not know whether the rain rule will be required.[15]

A very naïve rain rule would decide the winner of an abandoned match entirely by chance; for example, with the toss of a coin. This rule would have the equivalent correct prediction percentage as a rule that called every abandoned game a no-result, counting each no-result as half a correct prediction. This is our baseline measure. In order for a rain rule to have any credibility whatsoever it must predict the correct winner more often than this baseline. In

[15] This approach also keeps our series smooth, since if we were to use the alternative method of ignoring games already decided in the analysis, our sample size would be substantially reduced in the latter overs of the innings.

order to introduce the way that the CPP will be displayed throughout this chapter, we show the baseline in Figure 7.1.

Figure 7.1: The baseline CPP



7.3 The Average Run Rate (ARR) rule

The early methods of target adjustment in the event of rain were basic to say the least. The ARR rule simply asked the team batting second (Team 2) to score at a faster rate than the team batting first (Team 1), regardless of the number of overs available to each team. Let T_{ARR} be Team 2's revised target, S_1 be Team 1's score and O_1 and O_2 be the number of overs available to Team 1 and 2, respectively. The revised target score is as follows:

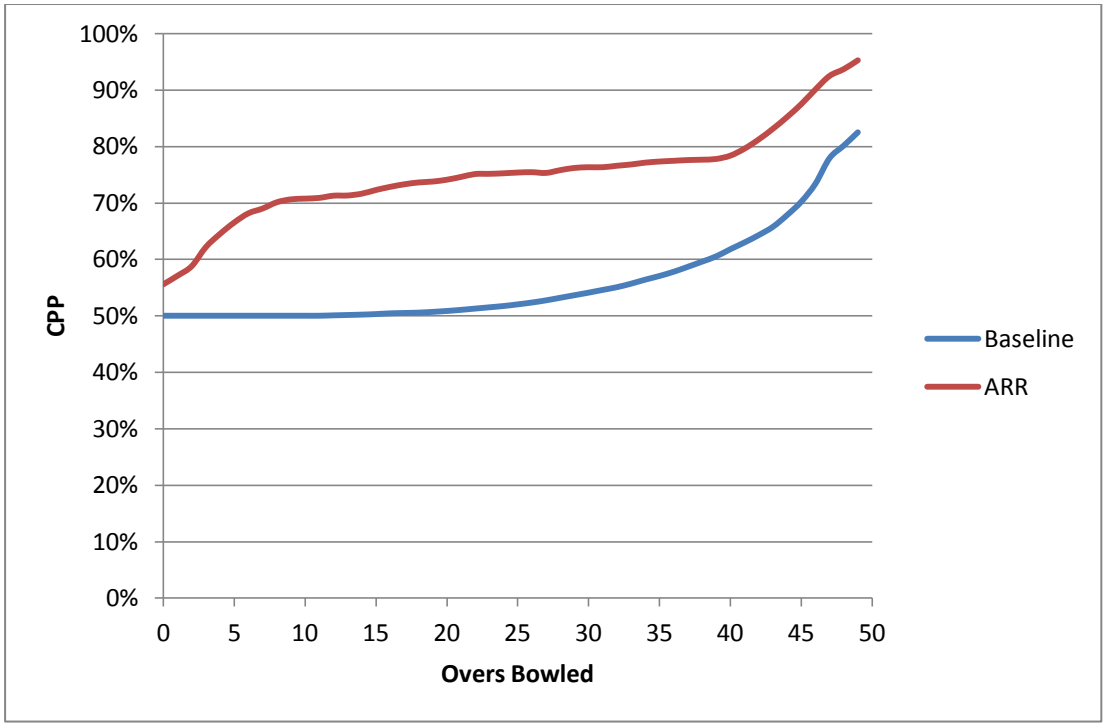
$$T_{ARR} = \frac{O_2}{O_1} \cdot S_1 + 1$$

The ARR rule has severe limitations. A batting team has two scarce resources – overs remaining and wickets in hand. The ARR rule reduces the target in proportion to overs lost but takes no account of the fact that an interruption does not affect Team 2's wickets in hand. Team 2 could bat very aggressively, scoring at a faster rate than Team 1 but losing nine of their ten wickets while being a long way short of the target when the rain comes. In an uninterrupted match, they would be very unlikely to win, but under the ARR rule they would be declared the winner by virtue of their faster scoring rate. This rule also ignores the timing of the interruption. Given a particular number of overs lost, it is clearly easier for Team 2 to plan their chase if the interruption occurs during the innings break than if their innings is unexpectedly cut short and the match is abandoned.

In Figure 7.2 we show the CPP of the ARR rule.[16] It is clear that the ARR rule, despite its shortcomings, is a substantial improvement on the baseline rule of deciding the winner by chance.

[16] In this and all subsequent CPP graphs we smooth the data using a 5-over centred moving average.

Figure 7.2: CPP – adding the ARR rule



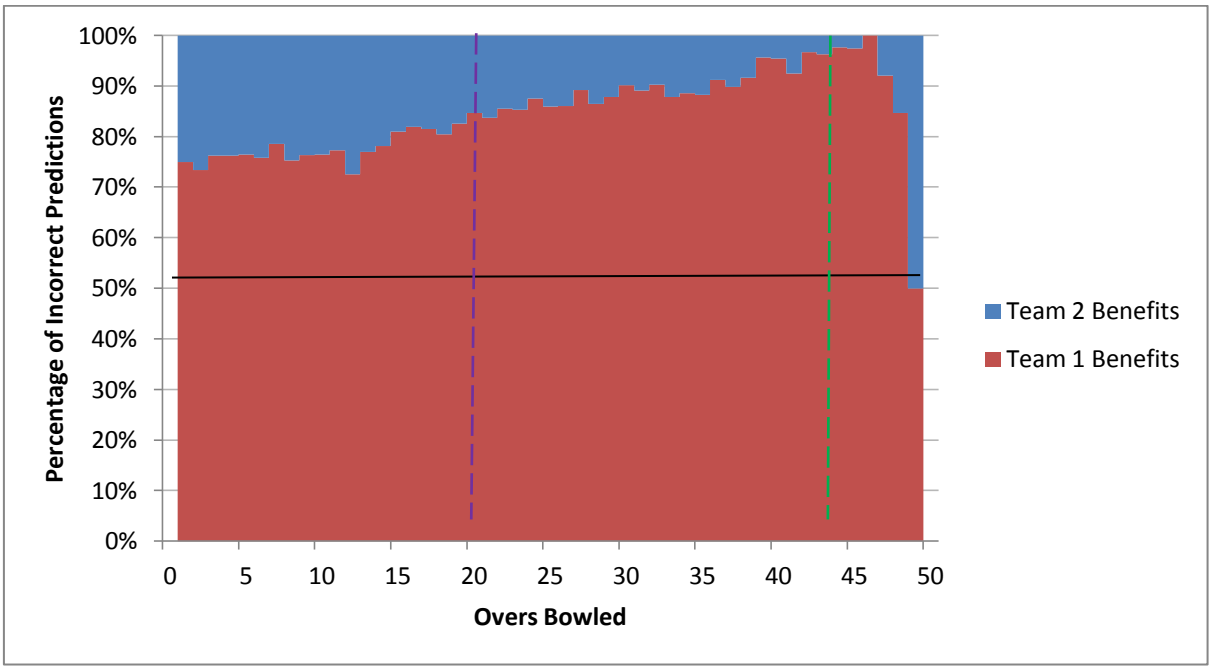
No rule is going to predict the correct winner 100% of the time. Cricket would be a very boring game if that were the case as it would mean the actions of players would be entirely predictable. The CPP, however, does provide a way of comparing two rules in terms of how well they used the match-specific information available prior to the interruption. It is not only the CPP that is important. We also consider those matches where the winner was incorrectly predicted by the ARR rule and look at which team would have benefited from the incorrect prediction - that is, which team would be incorrectly awarded the win in each case of abandonment. Figure 7.3 displays the percentage of the incorrectly predicted matches that would be awarded to each team.[17] It is clear that, aside from in the last three overs (where there is a very small sample size of matches still alive), Team 1 has a very large advantage under the

[17] In this and all future charts displaying the bias of a particular rule, we include a black line to show the theoretical line of fairness, a purple line at the 20-over mark to show the current minimum number of overs that constitute a match and a green line, to the right of which we have fewer than 50 incorrectly decided matches.

ARR rule in abandoned matches. This is because teams generally prefer to start their chase relatively conservatively, getting behind the required run rate but keeping wickets in hand with which to pick up the scoring rate towards the end of the innings. A surprise abandonment of the match is likely to find them short of their revised target.

The type of interruption least like an abandoned match is where the interruption occurs during the innings break. Logically, the advantage would be reversed in this situation as Team 2 would need to score at the same rate as Team 1, with all their wickets in hand, for fewer overs. We suppose that, for a given length of interruption, the fewer overs remaining for Team 2 at the resumption, the greater the advantage to Team 1.

Figure 7.3: Bias under the ARR rule



7.4 The Most Productive Overs (MPO) rule

The MPO rule involves the overs faced by Team 1 being ordered from the smallest number of runs scored to the largest. The reduced number of overs available to Team 2 is accounted for simply by removing Team 1's least productive overs, in terms of run scoring, from the target. That is:

$$T_{MPO} = \sum_{m=1}^{O_2} P_m + 1$$

Where P_m is the runs scored from the m^{th} highest-scoring over in Team 1's innings.

In Figure 7.4 we show that the MPO rule does not perform much better than simply tossing a coin to determine the winner. In fact, we show in Figure 7.5 that the MPO rule performs far worse than the baseline in terms of fairness, as when it makes a mistake it almost exclusively does so by being too quick to award the win to Team 1. The MPO method clearly advantages Team 1 as it only counts the overs of the first innings where they performed the best. Ignoring whether a predicted winner was correct or not and just focusing on the prediction itself, the MPO rule predicts Team 1 as the winner 97.9% of the time in an abandoned match**.** Team 2 would have more of a chance in a revised target situation where the match was resumed, but the size of the bias towards Team 1 in an abandonment situation makes is quite unbelievable that this method was used in important international matches. Indeed this method affected the 1992 World Cup, where South Africa's semi-final winning equation went from

needing 22 runs from 13 balls before a rain interruption, to needing an (almost) impossible 21 runs from a single ball at the resumption.

Figure 7.4: CPP – adding the MPO rule

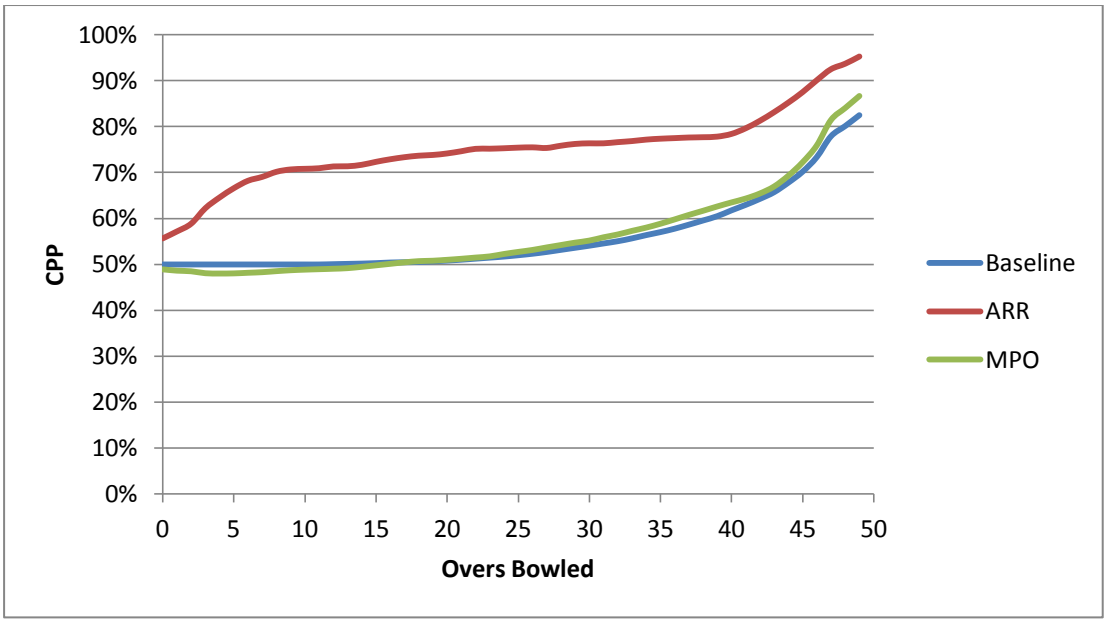
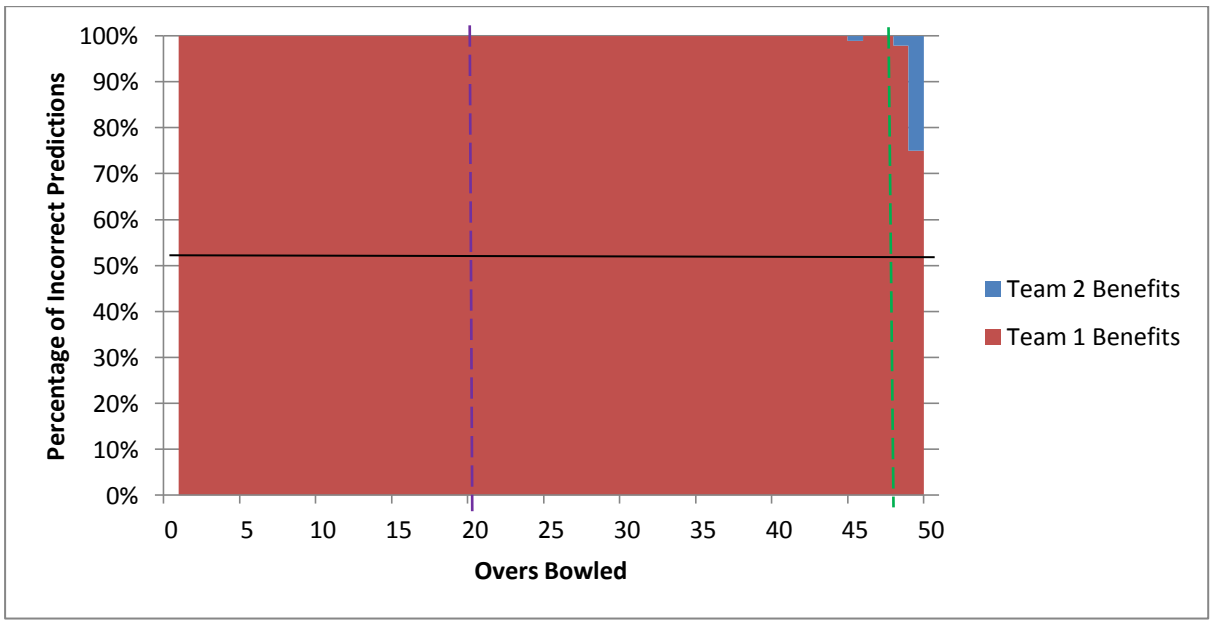


Figure 7.5: Bias under the MPO rule

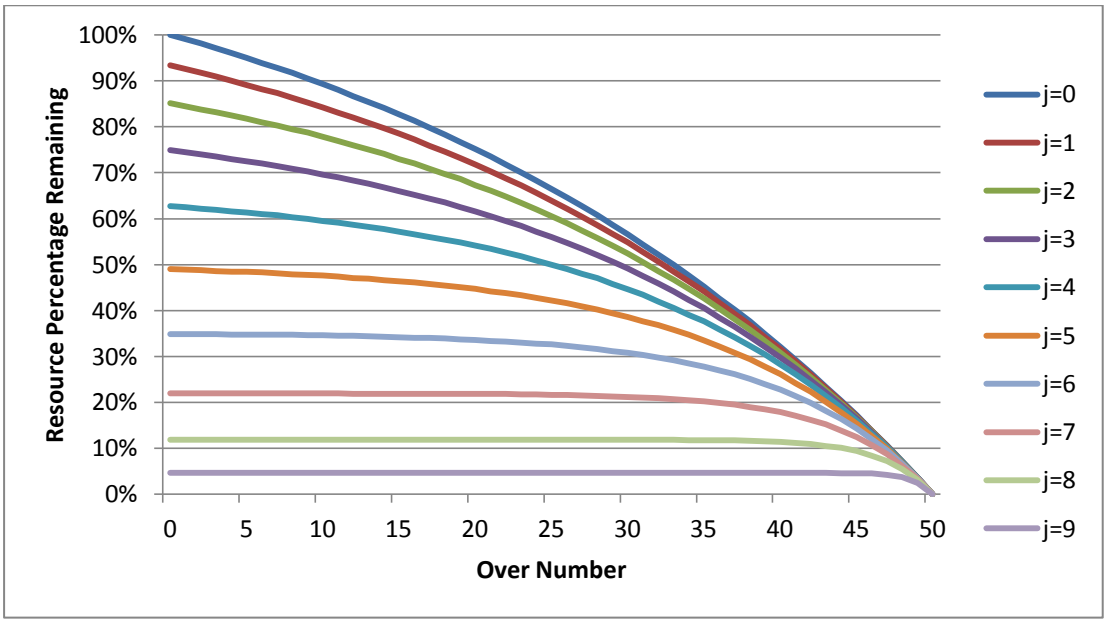


7.5 The Duckworth/Lewis (D/L) rule

Duckworth and Lewis (1998) proposed a method for calculating an appropriate target for the team batting second. Weather-affected ODI games have for more than a decade adopted the D/L rule. Like the ARR rule, the D/L rule is a resources-lost criterion however it has the significant advantage of taking both of a batting team's scarce resources, ball and wickets remaining, into account. To outline the basic idea of their method, consider the most common situation where two teams play a full length game. Each team has 100% of the resources (50 overs, ten wickets) of a full length game available to them. Team 2 simply has to beat Team 1's score, without adjustment. Now consider a game where Team 1 has an uninterrupted 50 over innings, but at some point in Team 2's innings it rains and ten overs are lost. Team 2 now only bats for a total of 40 overs. This clearly hurts their chances of beating Team 1's score so an adjustment is needed.

The D/L rule calculates the adjustment based on the ratio of resources available to Team 2 to the resources available to Team 1. If either innings is interrupted or abandoned, one or both of the teams lose some of their resources. The resources-lost depend on the number of overs and wickets remaining at the time of the interruption as shown in Figure 7.6 below. Citing commercial confidentiality, Duckworth and Lewis do not include the exact parameters of their model; however, they reveal that it is based on an exponential function for each wicket.

Figure 7.6: Duckworth/Lewis Resource Percentages



The D/L resources percentages are closely related to the V-functions (without conditions) that we calculated in Chapter 4. Duckworth and Lewis calculate the expected additional runs (which they call "average number of runs obtainable") and then divide each expected additional runs value by the initial expected runs at the start of an innings. This gives their resource percentages. The D/L model is based on fitting a model through data directly, as opposed to the dynamic programming approach that we used to create our V-functions in Chapter 5. We note that in this chapter we are mainly concerned with assessing the resources-lost criterion, rather than the precise method with which the resource percentages are calculated.

Additionally, Duckworth and Lewis have created a more sophisticated version of their model, which they call the professional edition, and this is now used in international cricket.

The basic concept is the same as the standard edition, but the resource percentages change depending on the first-innings score. Our assessment of their model is based on the publicly available standard edition. Defining the resource percentages available to Team 1 and Team 2 as R_1 and R_2, respectively, the formula for the Duckworth/Lewis target for Team 2 is[18]

$$T_{DL} = \frac{R_2}{R_1} \cdot S_1 + 1 .$$

This formula is essentially the same as the formula for the ARR rule, with the only difference being that the resource percentages are not constructed solely in terms of overs.

In Figure 7.7 we add the CPP for the D/L rule to our chart of the performance of each rule. It is clear that the D/L rule is a substantial improvement on the next-best ARR rule. Furthermore, Figure 7.8 does not suggest that the D/L rule substantially advantages one team or the other, in the case of abandoned matches, particularly considering the region between the green and yellow dashed lines – that is, after the minimum 20-overs have been bowled and before the data get thin.

[18] The formula is slightly different when Team 2 has a greater resource percentage than Team 1, a situation which would occur if Team 1's innings is unexpectedly cut short but Team 2 has full knowledge at the start of their innings of how many overs they have.

Figure 7.7: CPP – adding the D/L rule

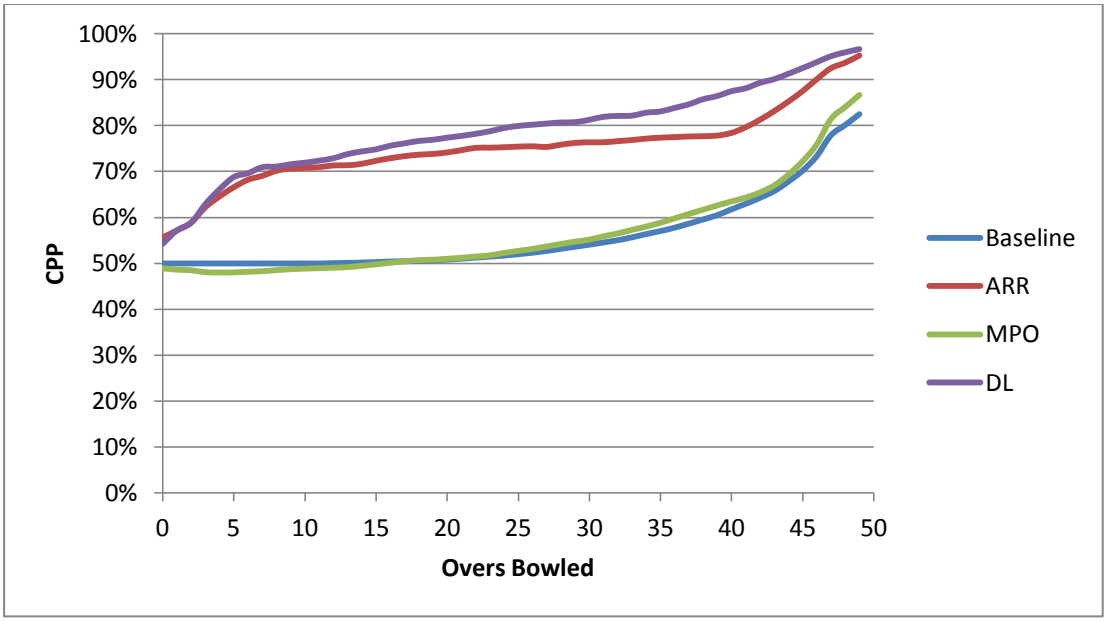
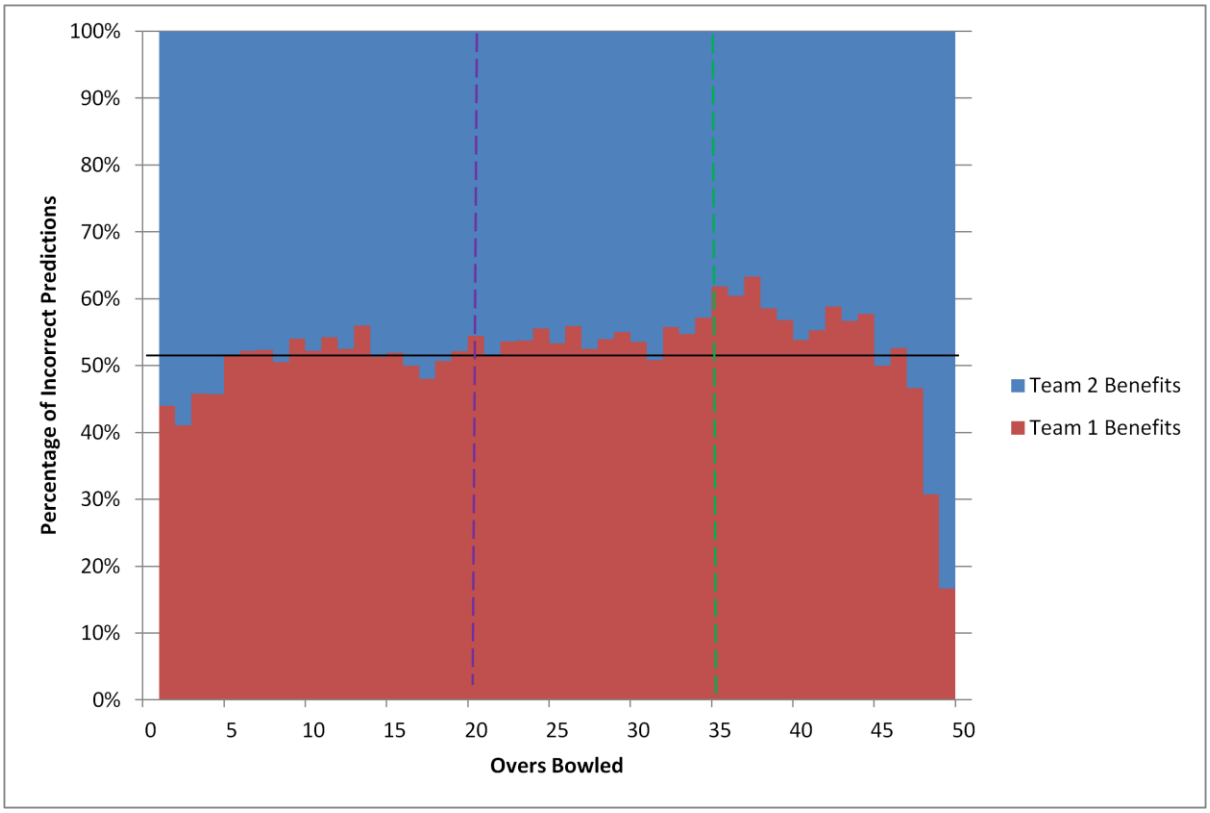


Figure 7.8: Bias under the D/L rule



7.6 A probability-maintenance criterion

Both the D/L and ARR rules use a "resources-lost" criterion in calculating a revised target score. The basic idea of this criterion is to assess the percentage of a team's resources that were lost over the course of the interruption or, in the case of an abandonment, the resource percentage that a team had available to them at the time of the abandonment. The revised target score is the nearest integer greater than Team 1's score multiplied by the resource percentage that Team 2 had available in total. The D/L rule uses a much more sophisticated model in determining what these resource percentages are in any given situation, compared to the ARR rule.

An alternative criterion is one of "probability-maintenance". The idea behind this criterion is that the probability of Team 2 winning the match is calculated for all possible situations. If an interruption occurs, Team 2's probability of winning is noted and when play resumes the revised target is set as the nearest integer greater than the target that gives Team 2 the same probability of winning after the interruption as they did before it. Thus the probability of winning is maintained across the interruption. In the event of an abandonment, the winner is declared to be the team with the greater probability of winning when play is stopped (a tie occurs if the probability of winning is exactly 50%.)

Preston and Thomas (2002) create a rule that uses a probability-maintenance criterion. They calculate their probabilities by assuming a model of run scoring and wickets, solving it for the optimal strategy, simulating 10000 games assuming that strategy and matching various percentiles of simulated scores to the distribution of first-innings scores in their observed data.

They repeated this process choosing a variety of values for the parameters of their model and chose the model that best matched their data.

Carter and Guthrie (2004) build on the work of Preston and Thomas (2002), arguing that the Duckworth-Lewis method is unfair in some situations. They use a dynamic programming approach to determine their probability of winning model, using the data from the 1999 World Cup in England.

We believe that a dynamic programming approach to the construction of the model is likely to result in a more accurate model, particularly in rare situations where there is little actual match data. Accordingly, we extend the work of Carter and Guthrie by creating a model of the second innings and include two significant additional variables: the run rate required and the ground conditions existing on the day of the match. We build our model both with and without the ground conditions variable. There are two main reasons for this: So that the difference between our rule and the incumbent D/L rule can be seen without the added complication of the newly-created conditions variable and so that the impact of including the conditions variable can be assessed.

7.7 The dynamic programme

In order to construct our new probability model, we set up a dynamic programme with which to estimate the probability of winning for Team 2, in any possible game situation. Our dynamic programme uses the same explanatory variables as the first-innings dynamic programme introduced in Chapter 5. In addition, we define a new state variable, the number of further runs required by Team 2 to win.

Let k be the number of further runs required, $k \in \{1, 2, ..., \infty\}$

The objective function in the second-innings model is the probability of Team 2 winning.

Let $\pi(i, j, k, \chi)$ be the probability of winning given i, j,k and χ. $\pi \in [0,1]$

The Bellman Equation is

$$\pi(i, j, k, \chi) = (1-\gamma_{ijk\chi}) \begin{bmatrix} (1-\lambda)\left(\sum_{r=0}^{6} \pi\left(i+1, j, k-r_{ijk\chi}, \chi\right)\right) \\ +\lambda_{ijk\chi}\pi\left(i+1, j+1, k, \chi\right) \end{bmatrix} + \gamma\left[\sum_{\omega=1}^{7} \omega_{ijk\chi}\pi(i, j, k-\omega, \chi)\right] \quad (25)$$

Equation (25) is a close approximation to the true process that is the second innings of an ODI but it is not an exact match. We assume that it is not possible for a batsman to be dismissed from a wide or a no-ball, whereas this is possible in cricket as one can be stumped or run out from a wide or run out from a no-ball. We further assume that it is not possible for runs to be scored from a ball upon which a batsman is dismissed, which can occur if a batsman is run out attempting two or more runs. We are, however, confident that these omissions affect the model only trivially.

In addition, it is possible that multiple wides or no-balls could occur sequentially, on a single legitimate value of *i*. Before estimating the model, we make the following adjustment to Equation (25).

Let $\psi_{ijk\chi}$ be a random variable indicating the sum of $r_{ijk\chi}$ and $\tau_{ijk\chi}$, the total runs from ball *i* whether they are from runs or extras $\psi_{ijk\chi} \in [0,13]$

Let $\Pr(\psi = \Psi) = \sum_{\Psi=R+\mathrm{T}} \sum \Pr(r_{ijk\chi} = R) \cdot \Pr(\tau_{ijk\chi} = \mathrm{T})$

The modified Bellman Equation is

$$\pi(i, j, k, \chi) = (1-\gamma_{ijk\chi}) \begin{bmatrix} (1-\lambda)\left(\sum_{\Psi=0}^{13} \pi\left(i+1, j, k-\psi, \chi\right) \Pr_{ijk\chi}(\psi = \Psi) \right) \\ +\lambda_{ijk\chi} \pi\left(i+1, j+1, k, \chi\right) \end{bmatrix}$$

The reason for this simplifying assumption is that the addition of the additional state variable for the number of runs required, *k*, already expands the state space by a factor of hundreds, when compared to the first-innings dynamic programme. Including the possibility of sequential wides or no-balls would add a significant additional complication to the estimation process since, unlike the first innings, we cannot simply take an expectation of the number of extras per legitimate ball, rather, we need to calculate each individual possibility separately. It is simply too computationally intensive to include the possibility of this rare event in our modeling process.

The value of k is, in most cases, meaningless without the value of i. If we know that a team needs 150 more runs to win, we have no idea how difficult its task is without knowing how many balls there are remaining. If they have all 300 balls in which to score the 150 runs, it is an easy task, but it is much more difficult if they only have 100 balls remaining. For this reason, we transform our state variable k into k^*, a variable representing the required run rate.

$$k^* = \frac{k}{301 - i}$$

Note that the state variable is still k - the use of k^* is restricted to the estimation process of r, λ, ω and τ.

Before we can proceed with the estimation of the dynamic programme, we need to filter our data set. The presence of a target score causes teams and players to behave differently to what they would do in the first innings. The first-innings goal is always to get as many runs as possible, regardless of the current situation. In the second innings, however, there are some game situations where it is obvious to all which team is going to win. When Team 2 has no realistic chance of winning the match, it is relatively common for its remaining batsmen to simply focus on spending some time batting, with no regard for the run rate required, which is likely to be high if the situation for Team 2 is hopeless. In the opposite scenario, when Team 2 is clearly going to chase the runs easily, they sometimes get very aggressive with the bat in a bid to finish the game quickly and, presumably, go celebrate the victory. These two approaches are unlikely to be optimal strategies; however, since the game is almost certainly lost or won anyway, employing these strategies make very little difference to the outcome. Unfortunately, they do contaminate our data set as we are attempting to use the required run rate as an

explanatory variable. In most situations it would be reasonable to expect aggression from the batting team, the higher is the run rate required; therefore, strategies in almost-decided games that are counter-intuitive can add considerable noise.

After investigating our dataset and looking at the likely run rate required / innings ball combinations where the game is essentially decided, one way or the other, we filter our data set by including data only if

$$\begin{aligned} &0.6 \le k^* \le 1.95, && i \ge 240 \\ &0.5 \le k^* \le 1.65, && 120 \le i \le 239 \\ &k^* \le 1.35, && i \le 119 \end{aligned}$$

We note that situations outside this range do not by any means make it impossible for either team to win the match; this filter is simply applied to ensure that our models are estimated from data where we are very confident that both teams are attempting to win the match. Having said that, we do apply some limits to the state space over which we estimate our dynamic programme for faster computation. We use the same principle to eliminate some states from the state space as we use to filter our data set, but our cut-off values cover a far greater range as we need to be almost certain that the true probability of winning would be approximately zero in these situations. The criteria for exclusion from the state space is

$$\begin{aligned} &8 \le k^*, && i \ge 289 \\ &3 \le k^*, && 241 \le i \le 288 \\ &2 \le k^*, && 181 \le i \le 240 \\ &1.75 \le k^*, && 121 \le i \le 180 \\ &1.5 \le k^*, && i \le 120 \end{aligned}$$

These limits ensure that we will not use valuable computing power estimating the probability of winning in states where Team 2 has an approximately zero probability of winning.

7.7.1 The regression equations

As we did for the first-innings model, we construct Probit and Ordered Probit models for $r, \lambda, \gamma \text{ and } \tau$. In selecting the variables and interaction terms to include, we consider p-values, coefficient size, cricket knowledge and consistency. This last factor involves giving greater consideration to a variable that appears significantly in, for example, the wickets regressions but not the runs regressions. In the tables in this chapter we show the model coefficients for the model with and without the conditions variable.

We decide to allow our second-innings model slightly more flexibility in shape than the first-innings model. This is because it was fairly obvious in the first innings that the scoring rate should increase as the innings progresses, for a given number of wickets lost, but this is less clear in the second innings as we need to consider the impact of runs required, which should be the most important variable for strategy. We create two new variables involving i and re-define i_3 from the first-innings model in order to order the variables sensibly.

Let $i_2 = \max(0, i - 48)$

Let $i_3 = \max(0, i - 180)$

Let $i_4 = \max(0, i - 240)$

Let $i_5 = \max(0, i - 270)$

We define the par score as the first-innings score which is closest to giving both teams an equal chance of winning. A simple Probit regression of the result on the first-innings score shows that there is a rather large difference in the par score in the power-play era, compared with the non-power-play era. These par scores are 248.2 and 265.3, respectively. With that in mind, we include the power-play variable in all our regressions, interacting it with other variables where appropriate. Interacting *j* with *i* proved relatively unimportant in the second innings, presumably because the required run rate is the most significant factor for chasing teams to have in mind. Additionally, there appears to be very little difference in runs or wicket patterns when a team is zero, one or two wickets down in the "without restrictions" period; therefore, these dummy variables are excluded.

The regression coefficients for our runs models are given in Appendix D. As expected, the new variable run rate required, introduced for the second-innings model, is an important inclusion. We see that the inclusion of the conditions variable has not changed the other coefficients substantially, although it appears that the required rate has a slightly smaller effect when conditions are taken into account. The coefficient of the conditions variable is positive, indicating that the scoring rate is higher in easier batting conditions.

Appendix D contains the coefficients and p-values for the wickets model. It is interesting that the interaction between the power-play variable and the required rate is significant and positive, indicating that chasing teams have responded better to the required rate in recent times. In contrast to the runs models, in the wickets models the inclusion of conditions

results in an increase in the significance of the required rate. The negative coefficients on conditions indicate that in better batting conditions, the probability of a wicket is reduced, as we would expect.

The coefficients and p-values for the probability of a wide or no-ball, and the probability of each number of runs from a wide or no-ball, are given in Appendix D. We see that the conditions do not substantially affect the probability of bowling a wide or no-ball, but good conditions do increase the number of runs that are scored of such deliveries on average.

7.7.2 The results of the dynamic programme

The dynamic programme is solved by backward induction after the calculation of the regression coefficients. If our dynamic programme is a good fit to the data, the initial par score (where $i=1, j=1$) implied by the version of the dynamic programme without conditions should be equal to the par score estimated by our Probit model regression result on first-innings score. The Probit model suggests a par score of 248.2 in the non-power-play era and 265.3 in the power-play era. Our dynamic programme suggests par scores of 249.5 and 263.0, respectively. This is a very close fit.

In the case of the conditions model, the par score should equal the value of conditions plus the performance advantage, as described in Chapter 4. These advantages are 2.0 in the non-power-play era and 5.8 in the power-play era. Our dynamic programme suggests the par scores shown in Table 7.1. We see that our model provides a reasonable, but not perfect fit. We tried fitting many different models and all resulted in the same effect: that the model

overestimates Team 2's chances of winning in poor conditions and underestimates it in good batting conditions. However, the most important result that we take from the analysis is the shape of the models and we correct our model by scaling it so that the par scores equal their theoretically-correct values.

Table 7.1: Par Scores from the conditions model

	Non-PP Era		**PP Era**	
χ	Par	Theoretical	Par	Theoretical
200	214.6	202.0	222.2	205.8
250	255.8	252.0	258.2	255.8
300	301.2	302.0	297.0	305.8

Our scaling method is to look at the probability of winning the match when the number of runs required at the start of the innings is equal to the theoretical par. We convert these probabilities to z-scores and we adjust the model by adding a constant to the z-scores so that the adjusted z-scores are equal to zero, for each combination of χ and era. These adjustments are shown in Table 7.2. We then convert the entire probability model to z-scores and add the same constant adjustment to all the situations for the given χ and era. The probability of winning at any point is then found from the cumulative standard normal distribution for the adjusted Z-score for the given *i, j,k* and χ .

Table 7.2: *Z*-score adjustments

Non-PP Era			
χ	Probability $(i=1, j=0, k=\chi, \chi)$	Z-score	Z-score Adjustment
200	0.5872	0.2203	-0.2203
250	0.5248	0.0622	-0.0622
300	0.4949	-0.0128	0.0128
PP Era			
χ	Probability $(i=1, j=0, k=\chi, \chi)$	Z-score	Z-score Adjustment
200	0.6191	0.3031	-0.3031
250	0.5176	0.0441	-0.0441
300	0.4384	-0.1550	0.1550

7.8 Assessing our without-conditions new probability (NP) rule

The NP rules uses the probabilities as calculated by the dynamic programme to make adjustments to the target score that preserve each team's probability of winning on either side of an interruption. In the case of an abandonment, the team with a greater than 50% chance of winning is declared the winner.

In Figure 7.9 we add the CPP for the NP rule to our chart of the performance of each rule. In Figure 7.10 we show the percentage of the incorrectly decided games that are awarded to each team. It is clear that the NP rule is a further improvement on the D/L rule. In particular, note that the NP rule outperforms the other rules most substantially early in the innings.

Figure 7.9: CPP – adding the NP rule

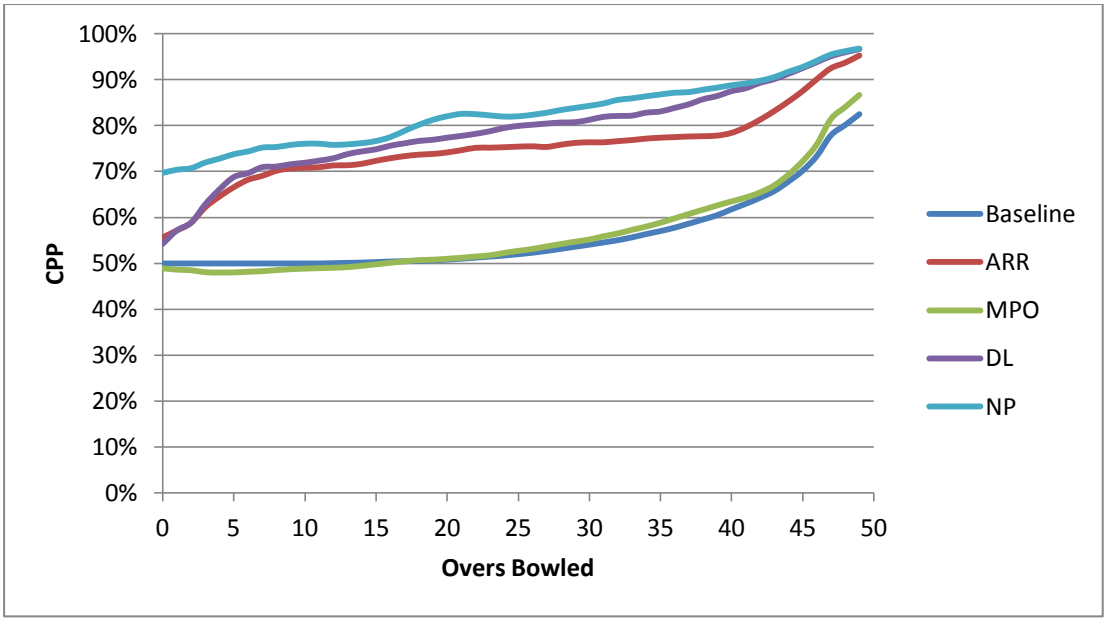
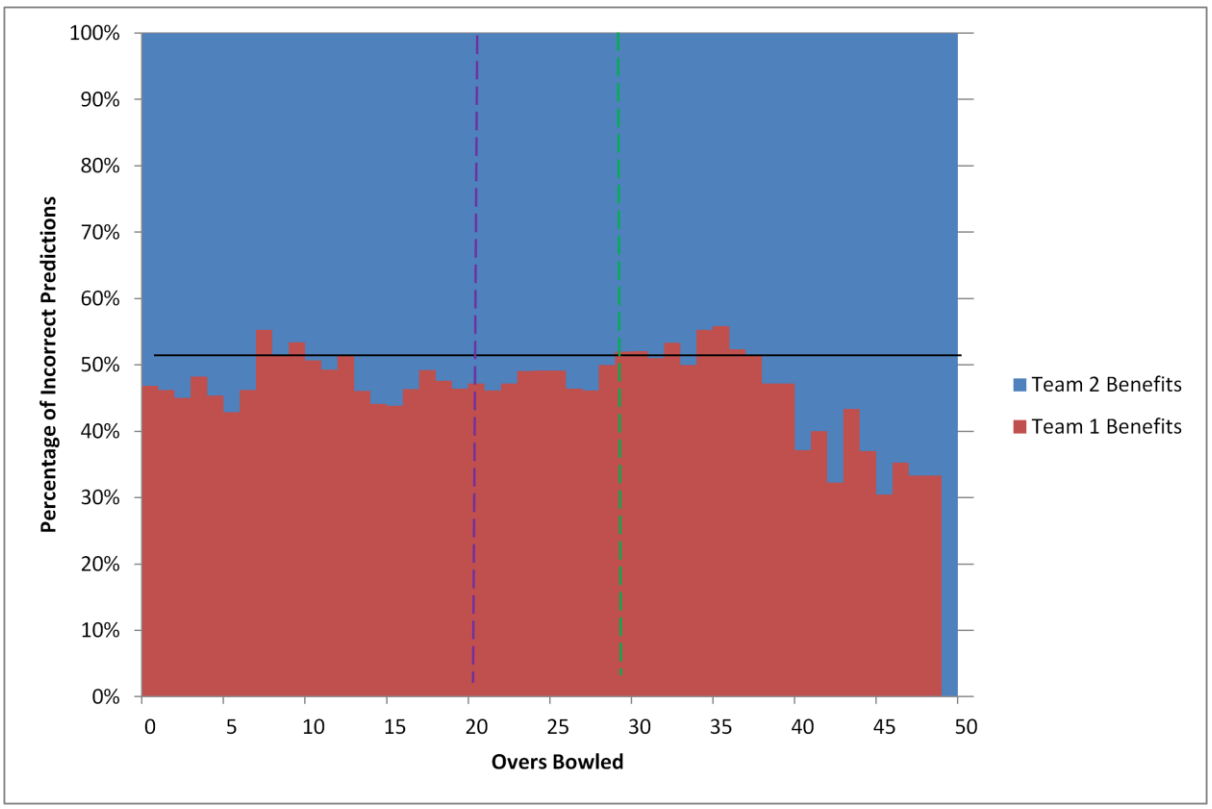


Figure 7.10: Bias under the NP rule



7.9 Assessing our with-conditions new probability (NP*) rule

Our ground conditions analysis in Chapter 4 provides us with the mean and variance of a conditional distribution for conditions, χ, given the score and result of the match. In applying our NP* rule, we randomly generate 100 values of χ for each start-of-over situation in our data set. Applying the probability model to these situations, we have 100 probabilities for each situation. We calculate the average probability from these 100 draws and use this average to predict the winner.

Calculating the probabilities is not completely straightforward as our computing power limitation means that we only have models for $\chi \in \{200, 250, 300\}$. Clearly, probabilities of winning exist for all other values of χ as well; however, it is not obvious what kind of function should be used to interpolate between these points. A simple linear scaling of the model is not a good method as this could lead to probabilities less than zero or greater than one. We select the option of fitting a Probit model where the latent variable is defined by a piecewise linear function. This ensures that the model fits our three data points and cannot result in a probability outside the range (0,1). The interpolation procedure is as follows.

Let Π_χ be the probability of winning in conditions worth χ
Let Z_χ be the z-score implied by the probability in conditions worth χ

1. Calculate Z_{200} and Z_{250}
2. Impose $Z_\chi = \alpha + \beta\chi$
3. Solve for α and β
4. Apply model to all $\chi \leq 250$
5. Repeat for Z_{250} and Z_{300}, and apply to all $\chi > 250$

The predictive ability of the NP* rule is shown in Figure 7.11. Clearly, adding the conditions variable into the model results in a more accurate model. Figure 7.12 shows the percentage of incorrect predictions that are made in favour of each team. By definition, the better the accuracy of a rule, the smaller the sample size from which to show any bias; however, it appears that the rule is treating the teams similarly.

Figure 7.11: CPP – adding the NP* rule

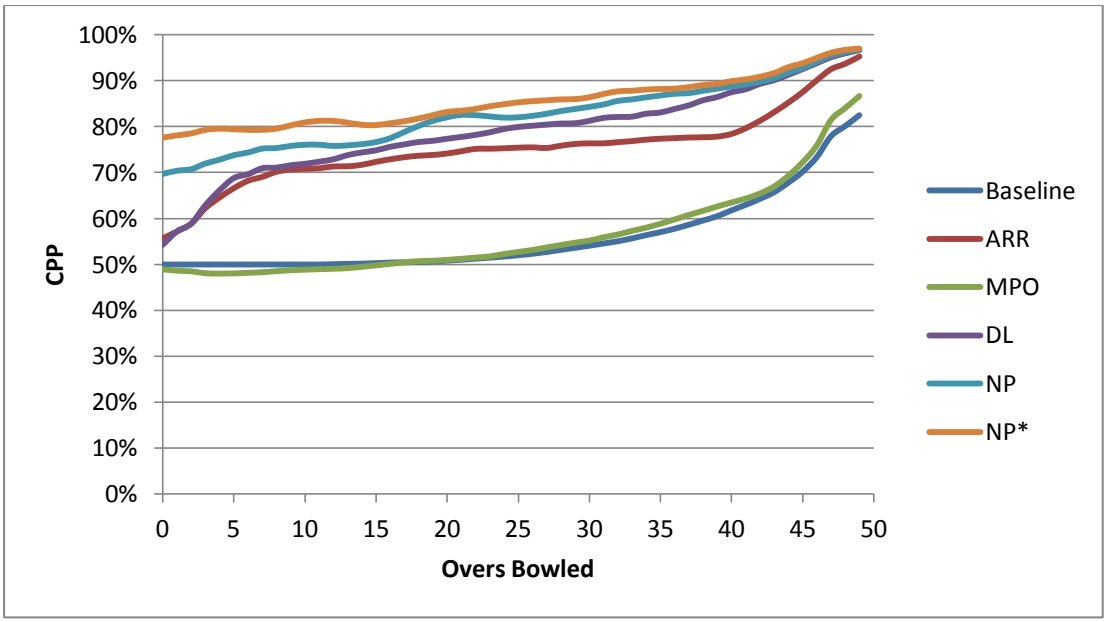
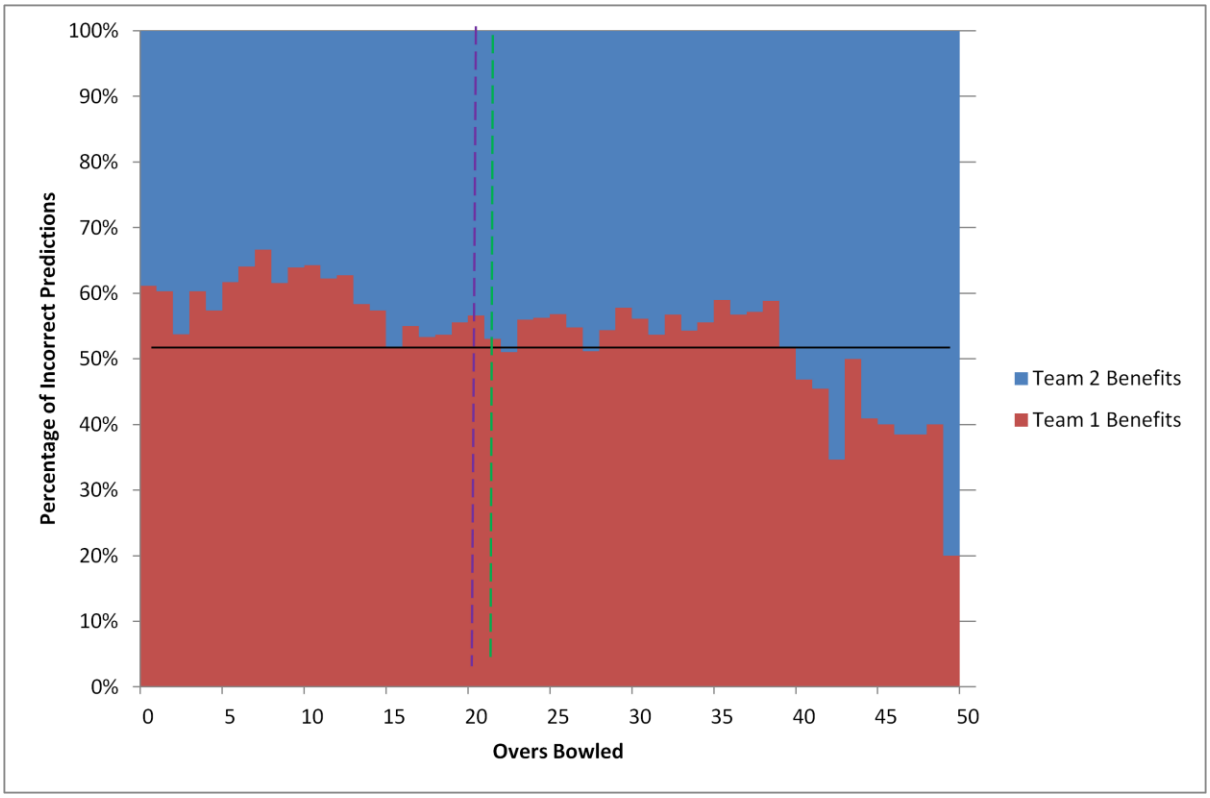


Figure 7.12: Bias under the NP* rule



A potential criticism of our NP* model is that the result of the game is used in estimating the mean and variance of the conditions distribution, effectively including the variable that the model is trying to measure as an explanatory variable. This raises the question of how an NP* rule could be implemented in practice, as our current estimation method requires knowledge of the result, which obviously is not available in an interrupted game. In practice, we would need a way of agreeing upon the value of conditions on the day of the match, before play takes place. One option would be to simply use the average first-innings total at that ground; however, one of the reasons for developing the conditions measure in Chapter 4 was that conditions vary substantially from match to match, even at the same ground. We propose that each captain and the match referee submit their estimate of the value of

conditions after inspecting the ground. The median estimate is then selected for the NP* rule for that match. The inclusion of the match referee and the use of the median, rather than the mean, prevents a captain with a weaker side submitting an outrageous estimate for conditions, hoping that the very unfair adjustment that would result goes in favour of his team.

Our conditions measure contains a substantial amount of uncertainty, in that it is a normal distribution with a mean and variance. Provided that this conditional distribution is representative of the true range of conditions that were likely to have been observed on the day of the match, we propose that using an expert opinion in the deciding of the conditions is only going to improve the predictive power of the model. While the estimation of the conditions variable is extremely useful for creating a variable to include in models and should be an unbiased estimator, cricket experts should be able to estimate a much more accurate point estimate for each match. Imposing a conditional distribution of conditions, simulating from this distribution and calculating the average probability is an inferior substitute for accurately estimating the value of the prevailing conditions, on the day of the match.

7.10 Three considerations when choosing a rain rule

There are three main considerations when constructing a rain rule. First, the criterion of fairness must be determined. Second, there is the choice of estimation method, of which two prominent ideas are direct estimation of scoring patterns from the data and indirect estimation via the calculation of transition probabilities in a dynamic programming framework. Third, the variables to be included in the modeling process need to be determined.

Like the Average Run Rate method, the Duckworth and Lewis adjustment uses a "resources-lost" criterion of fairness, where teams are compensated for the lost time with an adjustment reflecting the percentage of the target that teams would need to score, on average, from that lost time resource. The difference between the two methods arises from the difference in the way lost resources are measured, with the D/L rule taking into account the number of wickets remaining as well as the number of balls.

Carter and Guthrie discuss an example where Team 2 is playing very well and is obviously well ahead before an interruption and then later in the day further play is possible. In some situations the Duckworth-Lewis method determines that Team 2 is already ahead of their revised target score and therefore is declared the winner without any further play taking place. It must be the case, however, that Team 1 had some probability greater than zero of winning this game before the interruption but the adjustment hands them a loss with certainty.

We agree with the views of Carter and Guthrie on both the fairness of a probability-preserving adjustment rule and the dynamic programming approach. In our view, a team should have the same chance of winning after a rain interruption as they did before the rain. We illustrate the difference between the resources-lost criterion and the probability-maintenance criterion with a simple example. Consider a two-player sequential game where Player 1 tosses a fair coin 100 times, scoring one point for each "heads" and zero points for each "tails". Player 2 then plays the same game and wins if he has the higher score after his 100 tosses.

Let $S_{j,i} = \sum_{i=1}^{i} T_{j,i}$

Where $S_{j,i}$ is Player *j*'s score after *i* tosses and $T_{j,i}$ is Player *j*'s score from the i^{th} toss. Assume that Player 1 completes her 100 tosses and sets a score, but Player 2's turn is interrupted or has to be abandoned completely due to time constraints. Define toss *k* as the last toss that takes place before the interruption and toss *l* as the last toss that is lost to the interruption. If $l = 100$ then the match is considered to have been abandoned. Define $R_{k,l}$ as the resource percentage lost due to the interruption and $\pi_{2,i}$ as, after *i* tosses of his turn, Player 2's probability of winning. Finally, define Y_{RL} as Player 2's revised target under the resources-lost criterion and Y_{PM} as the revised target under the probability-maintenance criterion. We show the different effects of the two criteria for selected examples in Tables 7.3 to 7.5.

Table 7.3: $S_{1,100} = 50, k = 20, l = 80, S_{2,k} = 10$

$R_{20,80}$	60%
Y_{RL}	21
$\pi_{2,i}$	0.456
Y_{PM}	21

In this simple coin toss game, the resource percentage lost is simply the number of tosses lost divided by 100. In the example shown in Table 7.3, Player 2 is exactly on track to equal Player 1's score. In this situation, the revised targets given by the resources-lost criterion and the probability- maintenance criterion are identical.

Table 7.4: $S_{1,100} = 60, k = 20, l = 80, S_{2,k} = 15$

$R_{20,80}$	60%
Y_{RL}	25
$\pi_{2,i}$	0.109
Y_{PM}	29

Table 7.4 shows a situation where Player 1 has scored a very good score of 60 out of 100, but Player 2 has made an excellent start and is on 15 out of 20 at the time of the interruption. Player 2 still requires 46 heads out of his remaining 80 tosses and is clearly not the favourite. The probability-maintenance criterion recognises this and sets a target of 14 more heads out of the 20 tosses available after the interruption, while the resources-lost criterion sets a target of just ten more heads, clearly advantaging Player 2.

Table 7.5: $S_{1,100} = 25, k = 20, l = 80, S_{2,k} = 0$

$R_{20,80}$	60%
Y_{RL}	11
$\pi_{2,i}$	0.999
Y_{PM}	4

Table 7.5 shows a situation where Player 1 has scored a very bad score of just 25 heads out of 100 and Player 2 has started dreadfully, having not a single point on the board after his first 20 tosses. However, it is almost certain that Player 2 will be able to manage the 26 heads required from his remaining 80 tosses at the time of the interruption. The resources-lost

criterion fails to recognise this and sets a target of 11 heads from the remaining 20 tosses, compared to the four heads required by the probability-maintenance criterion.

The examples given show that the resources-lost criterion can cause very unfair revised targets to be set. The major problem with a resources-lost approach is that it assumes that the two players are equally likely to win after the first player's turn, regardless of what the first player scores, or equivalently, that Player 2 would have on average scored at exactly the required rate during the lost period of play, regardless of how easy or hard that required rate is. This results in a resources-lost criterion tending to, on average, dampen the advantage of the team who has played better so far in the match.

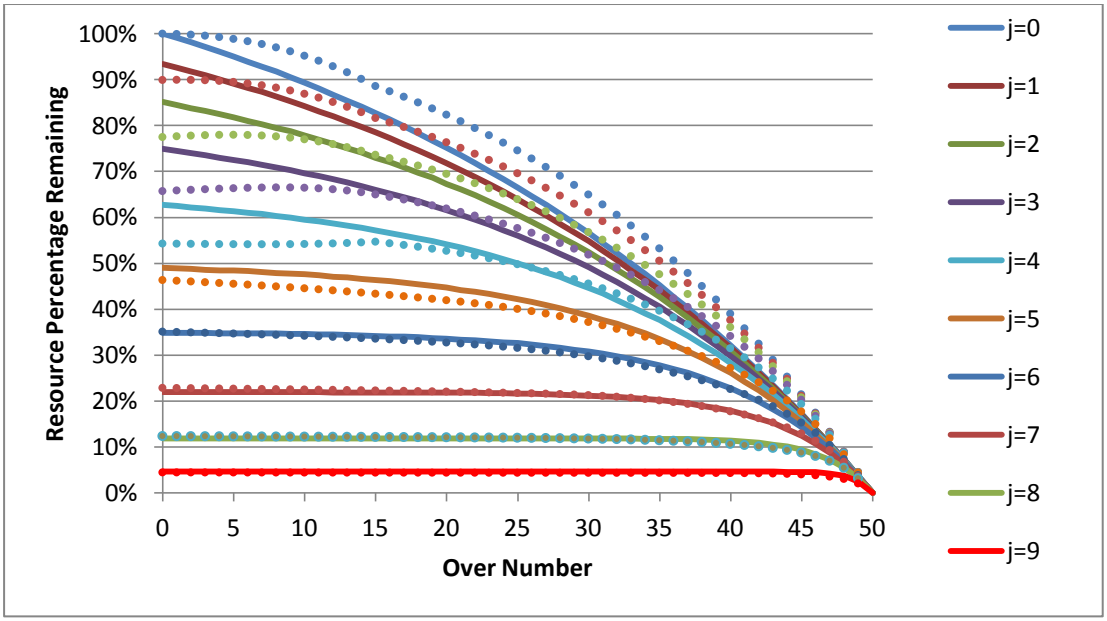
Estimating the probabilities of winning in a game of cricket is not as easy as in a coin-tossing game. In particular, the omission of a ground conditions variable in the estimation process leads to some circumstances where the resources-lost criterion can provide a fairer adjustment. Consider, for example, a situation where Team 1 achieves a very good score, but we know for sure that they batted only averagely, Team 2 bowled averagely and the very good score can be put entirely down to favourable batting conditions. In this case, starting the second innings as an even contest makes sense, as the implicit assumption of a resources-lost method is that all the variation in first-innings score is due to variations of conditions. We have shown in Chapter 4 that this is not the general case.

7.11 Decomposing the difference in predictive power

Our NP and NP* rules appear to outperform the alternative rules in predicting the results of abandoned matches. It is of concern, however, that our models were created from the data set that is subsequently being used to assess the predictive power of each model. This could lead to an over-fitting problem and may bias the results of the contest in favour of our model. In addition, we are interested in, as best as possible, isolating the difference in predictive power that is due to the choice of fairness criterion, the inclusion of the run rate required variable, and the inclusion of the conditions variable. These represent the major differences between the Duckworth/Lewis and Carter/Guthrie rules, the Carter/Guthrie and NP rules, and the NP and NP* rules, respectively.

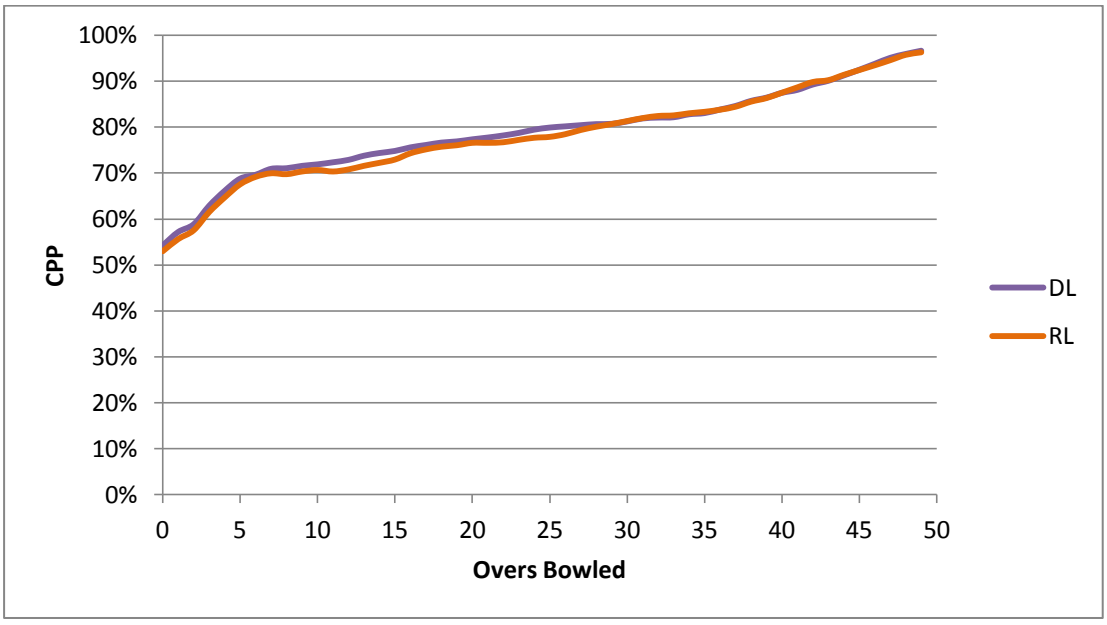
We approach this problem by constructing a DL-type model, from our second-innings data set. It is not an exact replica of the DL model as we do not know their exact formula; instead we use our first-innings dynamic programme (without conditions) to estimate the expected additional runs for any i and j. We then divide these expected additional runs values by the expected additional runs in the $(i=1, j=0)$ cell, in order to get the resource percentages. This does not give the DL-model, but it gives a model which has been created based on average scoring patterns, with both ball of innings and wickets lost taken into account. Crucially, this model uses a resources-lost criterion, identical to the DL. We call this new rule the RL rule and we plot the resources percentages given by the RL rules (as dotted lines) against the DL resource percentages in Figure 7.613. We note that there is substantial difference in the early part of the innings. This possibly can be put down to the dynamic programming approach of the RL curves compared to the assumed functional form approach of the DL curves.

Figure 7.13: DL versus RL resource percentages



In Figure 7.14 we show that there is very little difference between the predictive power of this rule and the DL rule. If anything, the DL rule does slightly better, which we speculate could be due to having a much larger sample size with which to estimate their model.

Figure 7.14: DL versus RL



Next, we create a CG-type model; that is, a rule which uses the probability-maintenance criterion, but does not include the run-rate-required variable. We simply re-run our NP model without this variable and we call the resulting model the CG~ rule. The predictive power of the GG~ rule is plotted alongside that of the NP rule in Figure 7.15.

There is very little difference between the CG~ and NP rules, suggesting that the run-rate-required variable has little impact. Upon closer inspection, this is not the case. The CG~ rule tends to underestimate the probability of the team which is currently losing making a comeback, when compared to the NP rule. We show this in Figure 7.16 for the situation at the start of the second innings. This makes intuitive sense as a team who is losing will, more likely than not, require a high run rate and the NP rule takes into consideration their increased urgency. Predicting a winner in an abandoned game simply assesses which team is winning at the time of the abandonment, and CG~ and NP are similar here. This does not, however, mean that the two rules would necessarily give similar revised targets in a match that resumes after the interruption.

Figure 7.15: NP versus CG~

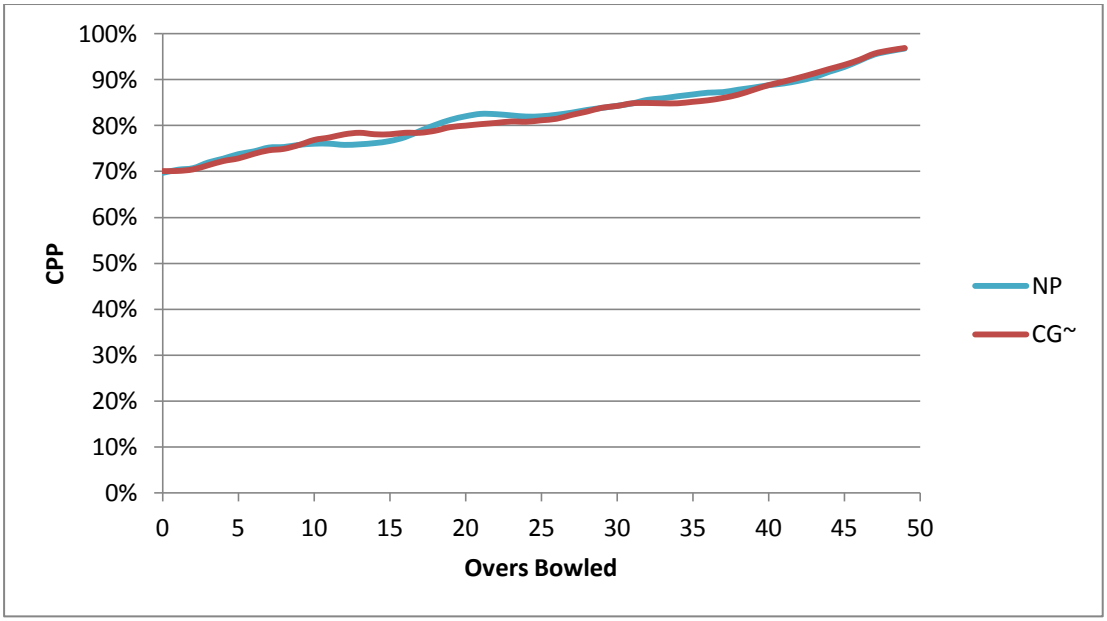


Figure 7.16: Probabilities for *i=1, j=0, pp=0*

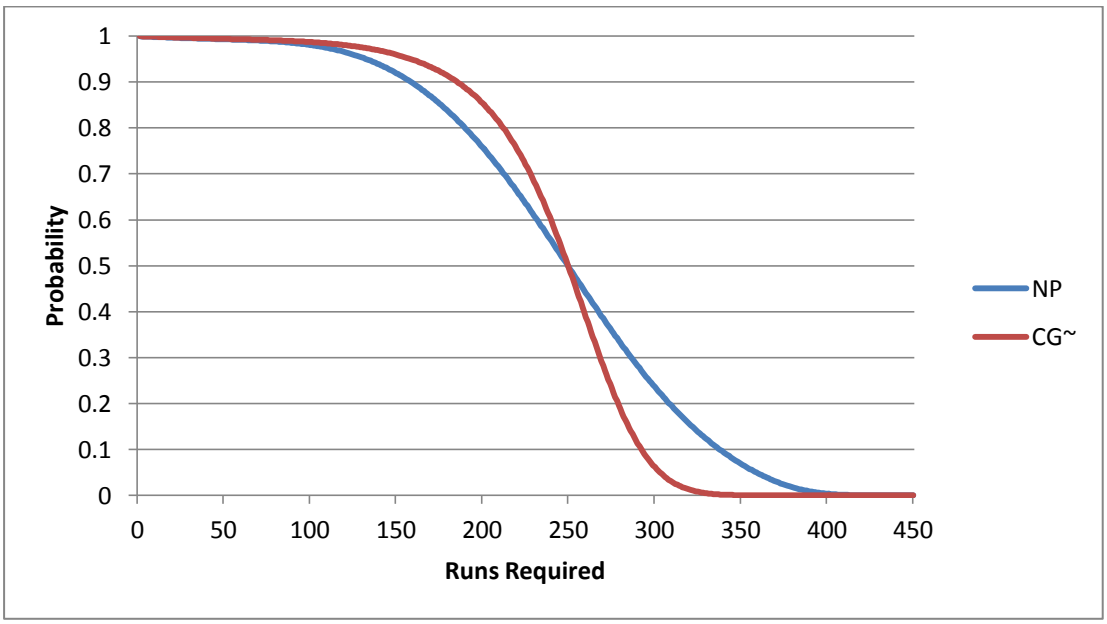
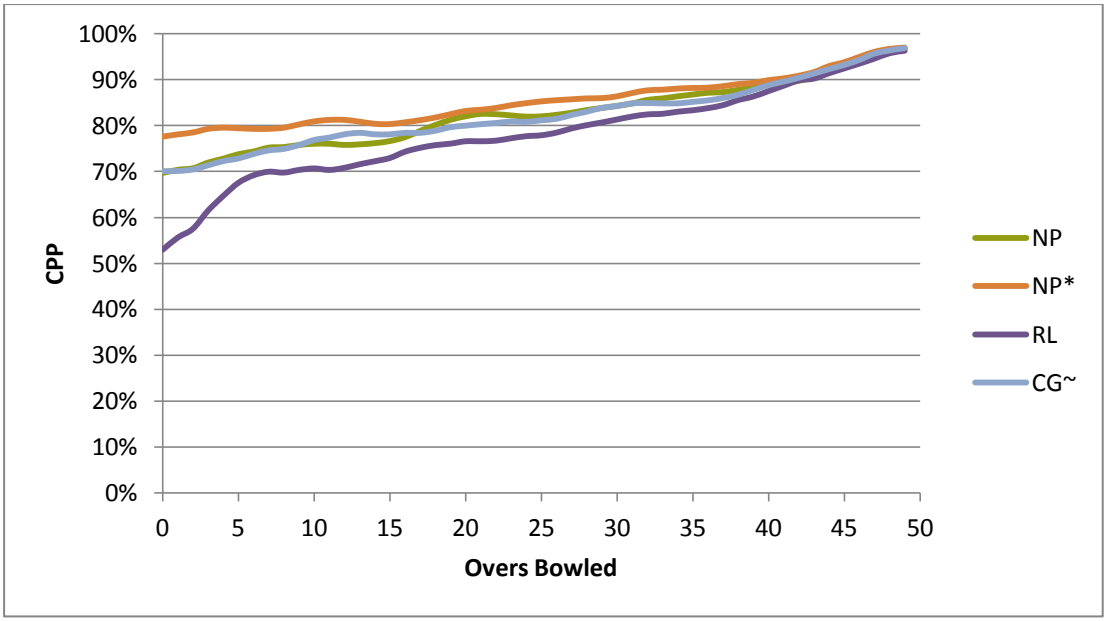


Figure 7.17 shows the difference between the predictive power of the RL rule, the CG~ rule, the NP rule and the NP* rule. It is clear that the largest difference is made by the use of a probability-maintenance criterion, compared to a resources-lost criterion. Including conditions is also a substantial improvement, while the incorporation of the run rate required variable makes little difference in terms of predictive power (although as previously noted it does make a difference to the revised targets that would be set in a resumed match).

Figure 7.17: All rules created from the same data set



7.12 Discussion of the different criterion in resumed matches

We have shown that there is significant variation in the performance of the rain rules that we assess in predicting the winner of abandoned games. It is difficult to apply this assessment to resumed matches as the strategy of the teams after the resumption would depend

substantially on the target that they are set. It is possible to make some general statements about the impact that we expect each rule to have in resumed matches.

An accurately estimated probability model allows the probability-maintenance criterion to be applied in such a way that the revised target advantages neither team, as each team would have the same probability of winning after the interruption as they did prior to it. Assuming that the NP* model, as the most accurate predictor of winners in abandoned games, is an accurate model, we can be sure that the ARR and D/L rules, both resources-lost criterion, will sometimes set unfair targets. However the exact impact of the use of an unfair method is complicated and varies with the particular situation of the game.

We take our data set with abandonments and alter it to allow for matches to be resumed after a break. We use break lengths of 5, 10, 15, 20, 25 and 30 overs for each interruption, only including those that allow a resumption in play. We outline a brief comparison of the revised targets that would be set by the NP* method and those that would be set by the D/L method. Table 7.6 shows the summary statistics of the difference in these revised target scores as the D/L target minus the NP* target.

Table 7.6 D/L revised targets vs. NP* revised targets

Statistic	Value
Mean Difference	-12.6
Median Difference	-13
Minimum Difference	-85
Maximum Difference	+167

There are some very large differences in the revised targets. The difference of -85 came about in a game where Team 1 had scored 185 runs and Team 2 were at 133/1 after 15 overs

when the rain came. Thirty overs were lost and therefore just five overs remained upon the resumption of the match. The conditions were worth 279 runs and Team 2 had a probability of winning this match of 0.996. The D/L revised target is 72 runs, meaning that Team 2 were already 61 runs past the target. The NP* revised target is 157, meaning that Team 2 would have to score a further 24 runs from the last 30 balls, with nine wickets in hand. This is a very easy task that matches their extremely high probability of winning before the break. However, unlike under the D/L rule, Team 1 still has a slim glimmer of hope, as they did before the break.

At the other end of the scale, in one situation the D/L rule sets a target of 167 more runs than the NP* rule. This occurred when Team 1 had the very high score of 392 and Team 2 were struggling at 125/8 after 20 overs. There was almost no chance that Team 2 would win this game (the probability was 0.00001). After a 25 over break, the D/L rule makes sure they cannot win by asking them to score a further 258 runs from the last 30 balls. The NP* rule asks them to score a further 91 runs, which is almost impossible, but better reflects the position that each team was in before the break.

The mean and median difference indicate that overall the D/L rule favours Team 2 by setting revised targets that are too low. This seems consistent with the comments that captains tend to make upon winning the toss in a match where rain is forecast for later, as they usually indicate that they think they are better to be batting second in the event of the D/L rule being invoked.

Throughout the course of this chapter, we have shown that rain rules have certainly undergone substantial improvement since the days of the ARR rule and the simply awful MPO

rule. We have also shown, however, that it is possible to do substantially better than the DL rule, particularly by selecting probability-maintenance as the fairness criterion, rather than resources-lost. In addition, this chapter has contributed a method for comparing the predictive power of different rain rules. If this diagnostic had been employed prior to the adoption of the MPO rule, it would likely have never been seriously considered for use in actual games.

CHAPTER 8

Concluding remarks

We hope that, having reached the end of the thesis, the reader has an appreciation for the immense value that a variable for the ground conditions adds to any cricket analysis. By using this variable, research will not be subject to the criticism that the findings could simply be explained by variation of ground conditions over the course of the data set. Furthermore, we provide an insight into the strategic information that can be gained by plugging in different values of ground conditions into a model and developing strategies suitable for particular conditions.

The use of the cost of a wicket as a proxy for the risk taken by a batsman enables us to construct PPFs to represent the ability of individual batsman. We show that it is possible to separate the two components of a batsman's average performance in particular situations – natural ability and strategic nous. This is useful in many different ways. Estimates of the raw ability of batsmen might be used in identifying talented players while the strategic nous measures provide an indication of which players have a good feel for what is required in particular situations. It becomes possible to identify the batsmen who are currently performing close to their maximum ability and to work without those who could become better players by improving their tactical awareness. The PPFs would also useful for quantifying trade-offs such as identifying the impact of selecting an extra batsman or to address the issue of optimal batting orders.

We show that while the Duckworth/Lewis method is easily the best of the resources-lost methods used for target-adjustment in weather-affected matches, there is substantial room for improvement. By artificially terminating complete games we show that a probability-maintenance criterion outperforms the Duckworth/Lewis method in predicting actual results correctly. In addition, we show that the incorporation of a ground-conditions variable in the rain rule results in a further improvement to the predictive power of the probability-maintenance model.

List of References

Allsopp, P.E. and Clarke, S. R. (2004). Rating teams and analysing outcomes in one-day and test cricket. *Journal of the Royal Statistical Society. Series A (Statistics in Society)* **167(4):** 657-667.

Bailey, M. and Clarke, S. R. (2006). Predicting the match outcome in one day international cricket matches, while the game is in progress. *Journal of Sports Science and Medicine* **5:** 480-487.

Barr, G.D.I. and Kantor, B.S. (2004). A criterion for comparing and selecting batsmen in limited overs cricket. *Journal of the Operational Research Society* **55(12):** 1266-1274.

Brown, L.D., Cai, T.T, and Dasgupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science* **16(2):** 101-133.

Carter, M and Guthrie, G (2004). Cricket Interruptus: fairness and incentive in interrupted cricket matches. *Journal of the Operational Research Society* **55:** 822-829.

Carter, M and Guthrie, G (2005). Reply to the comments of Duckworth and Lewis. *Journal of the Operational Research Society* **56:** 1337-1341.

Clarke, S.R. and Norman, J.M (1999). To run or not? Some dynamic programming models in cricket. *Journal of the Operational Research Society* **50(5):** 536-545.

Clarke, S.R. and Norman, J.M (2003). Dynamic programming in cricket: choosing a night watchman. *Journal of the Operational Research Society* **54(8):** 838-845.

Clarke, S. R. (1988). Dynamic programming in one-day cricket – optimal scoring rates. *Journal of the Operational Research Society* **39(4):** 331-337.

Duckworth, F.C. and Lewis, A.J. (1998). A fair method for resetting the target in interrupted one-day cricket matches. *Journal of the Operational Research Society* **49(3):** 220-227.

Duckworth, F.C. and Lewis, A.J. (2004). A successful operations research intervention in one-day cricket. *Journal of the Operational Research Society* **55(7):** 749-759.

Duckworth, F.C. and Lewis, A.J. (2005). Comment on Carter M and Guthrie G (2004). Cricket Interruptus: fairness and incentive in limited overs cricket matches. *Journal of the Operational Research Society* **56:** 1333-1337.

Hirotsu, N and Wright, M (2003). Determining the best strategy for changing the configuration of a football team. *Journal of the Operational Research Society* **54:** 878-887.

Jayadevan, V (2002). A new method for the computation of target scores in interrupted, limited-over cricket matches. *Current Science* **83(5):** 577-586.

Klaassen, F.J.G.M. and Magnus, J.R. (2008). The efficiency of top agents: an analysis through service strategy in tennis, *Journal of Econometrics* **148**: 72–85.

Lewis, M. (2003), *Moneyball*, New York: W. W. Norton & Company.

Manage, A. B. W., Kumudu, M. and Kanchana, W. (2010). Receiver Operating Characteristic (ROC) curves for measuring the quality of decisions in cricket. *Journal of Quantitative Analysis in Sports* **6(2):** Article 8.

Norman, J.M. and Clarke, S.R. (2010). Optimal batting orders in cricket, *Journal of the Operational Research Society* **61:** 980-986.

Preston, I. and Thomas, J. (2000). Batting strategy in limited overs cricket. *Journal of the Royal Statistical Society: Series D (The Statistician)* **49:** 95-106.

Preston, I. and Thomas, J. (2002). Rain rules for limited overs cricket and probabilities of victory. *Journal of the Royal Statistical Society: Series D (The Statistician)* **51(2):** 189-202.

Romer, D. (2003). It's fourth down and what does the Bellman equation say? A dynamic-programming analysis of football strategy. NBER Working Paper No. 9024.

Appendix A: Numerical investigations into the split of performance and conditions

It is useful to test whether the distribution of scores, S_2, implied by the distribution function $J_\delta(S_2)$ is normal. Normality would enable the simple calculation of the mean and variance of the second-innings distribution as we could then determine $J_\delta(S)$ for two values of *S* and fit a straight line through the Z-scores implied by those two points. In order to test for normality we arbitrarily split the first-innings variance ($\sigma_S^2 = 3412.488$) as 60% due to performance and 40% due to conditions $\left(\delta = 60, \sigma_\rho^2 = 2047.493, \sigma_\chi^2 = 1364.995\right)$ and calculate $J_{60}(S)$ for all values of S in the interval (0, 500), which easily covers the range of observed scores. If $J_\delta(S)$ implies a normal distribution of S_2 then the Z-scores associated with each value of $J_\delta(S)$ will be linear in *S*. In Figure A.1 we plot $J_{60}(S)$ and in Figure A.2 we plot the Z-scores. A linear regression of Z on S results in the following:

$$Z = \alpha + \beta S$$

$$Z = -2.726433718 + 0.011206656S$$

$$R^2 = 1.00$$

It is apparent from the R^2 value of one that the distribution of S_2 implied by $J_{60}(S)$ is perfectly normal. We assume this result applies to all values of $(0 < \delta \leq 100)$. Note that $\delta = 0$ is a special case as when there is no variation in performance all matches will be tied; therefore, if we are

consistent with our earlier treatment of counting ties as both a win and a loss, each with a weight of 0.5, $J_0(S) = 0.5$, regardless of the value of *S*.

Figure A.1: The implied second-innings distribution

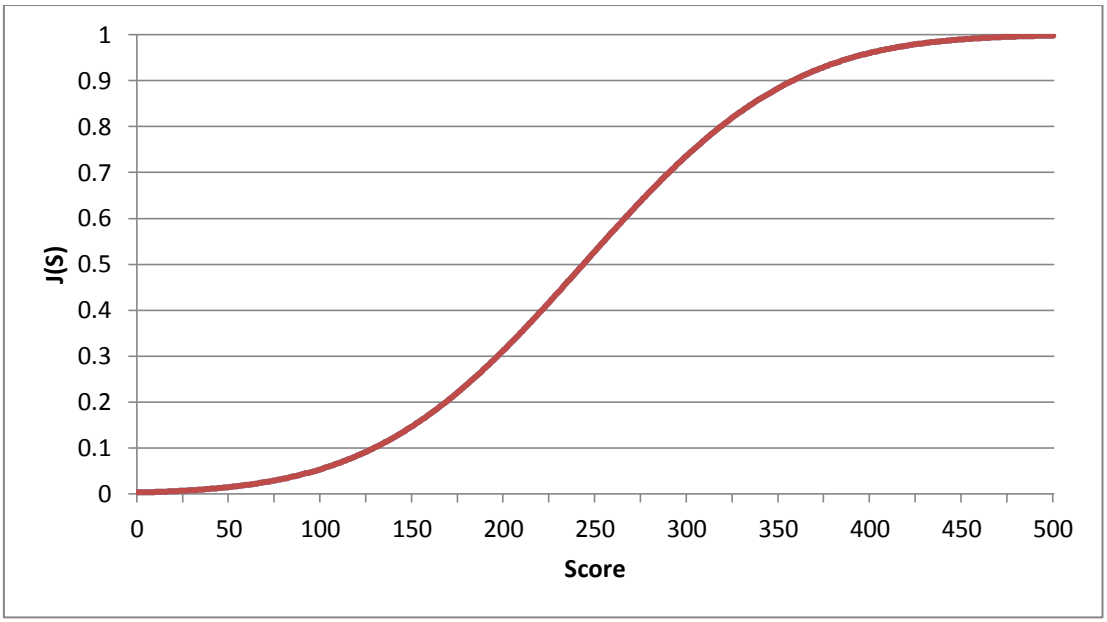
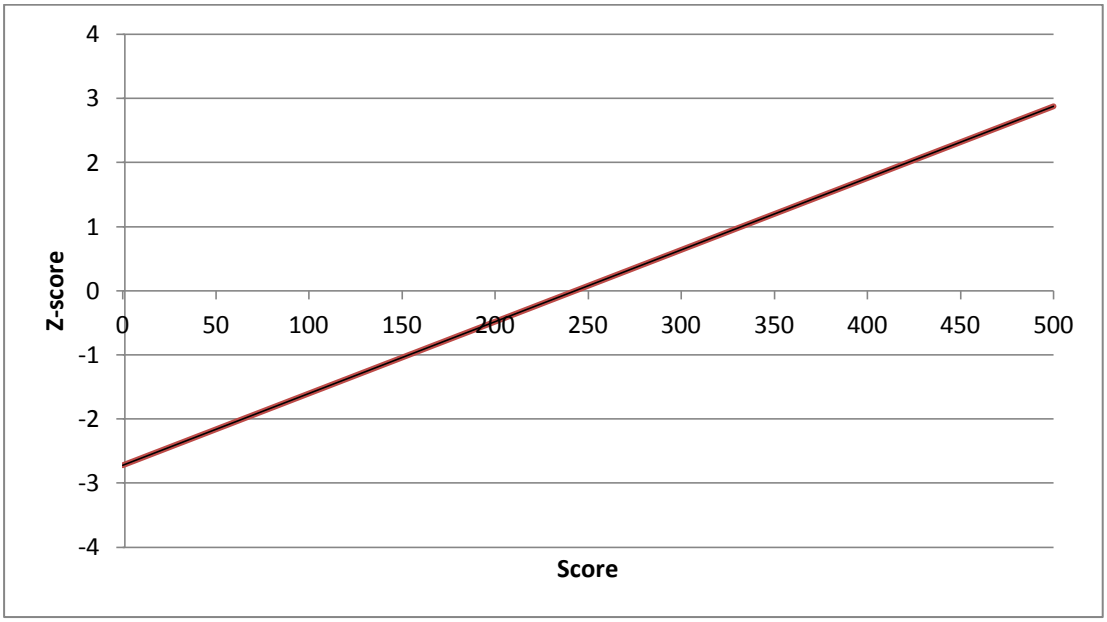


Figure A.2: Z-scores implied by $J_{60}(S)$



Given the normality of $J_\delta(S)$, we are able to estimate the mean of the implied second-innings distribution, S_2, using the function

$$J_\delta(S) = \Phi(Z_\delta)$$

where Φ is the cumulative distribution function of the standard normal distribution. Since *Z* is linear in *S*, we can write

$$Z_\delta = \alpha_\delta + \beta_\delta S \qquad (26)$$

The mean and variance of the second-innings distribution implied by Equation (26) are

$$\mu_{S_2,\delta} = \frac{-\alpha_\delta}{\beta_\delta} \tag{27}$$

$$\sigma^2_{S_2,\delta} = \frac{1}{\beta^2_\delta} \tag{28}$$

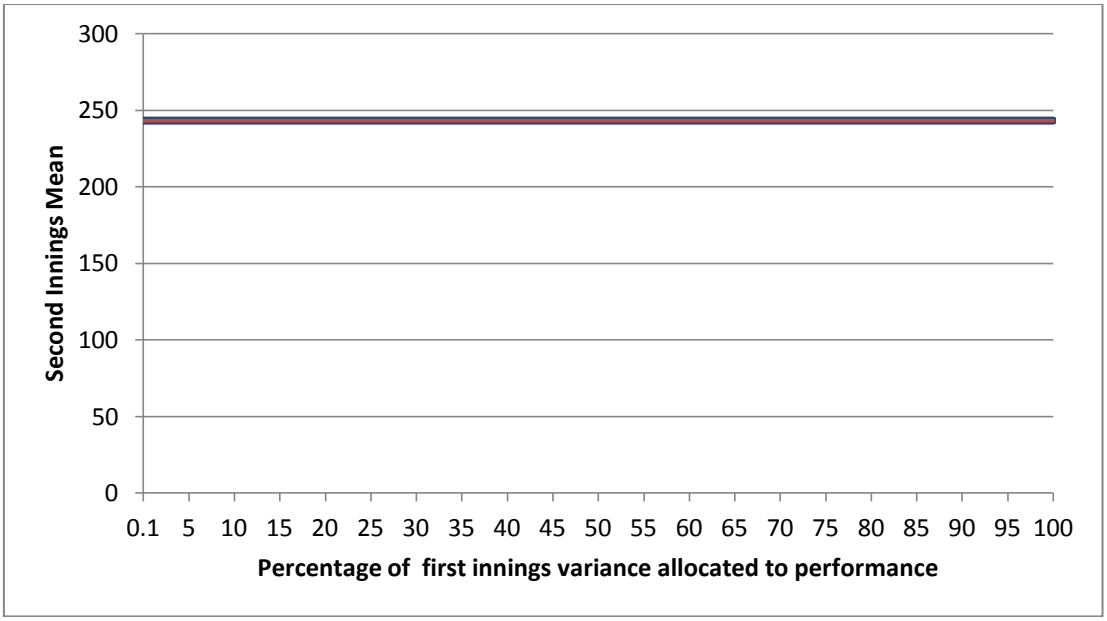
From our example of $\delta = 60$, the Probit regression shows that $\alpha_{60} = -2.726433718$ and $\beta_{60} = 0.011206656$. Plugging these values into Equations (27) and (28) results in $\mu_{S_2,60} = 243.287$ and $\sigma^2_{S_2,\delta} = 7692.472$. The mean, as expected, is equal to our observed first-innings mean (recall that we have ignored the second-innings advantage in this analysis to date), while the variance is greater than our observed second-innings variance of 5673.117. This tells us that we have allocated too much of the first-innings variance to conditions and not enough to performance, as the second-innings variance would get larger, the larger is the conditions variance.

We set up a macro in SAS to split the first-innings variance into performance variance and conditions variance 21 different ways (0.1%,[19] 5%, ..., 95%, 100% of the total variance is allocated to conditions) and calculated the implied second-innings variance in each case. We note that, while we calculated a regression equation in our example above, with *Z* being linear in *S* it is only necessary to calculate the value of *Z* for any two values of *S* in order to determine

[19] We know that a performance variance of zero will result in a second-innings variance of ∞, so we choose a performance variance very close to zero in order to demonstrate this result in a calculable way.

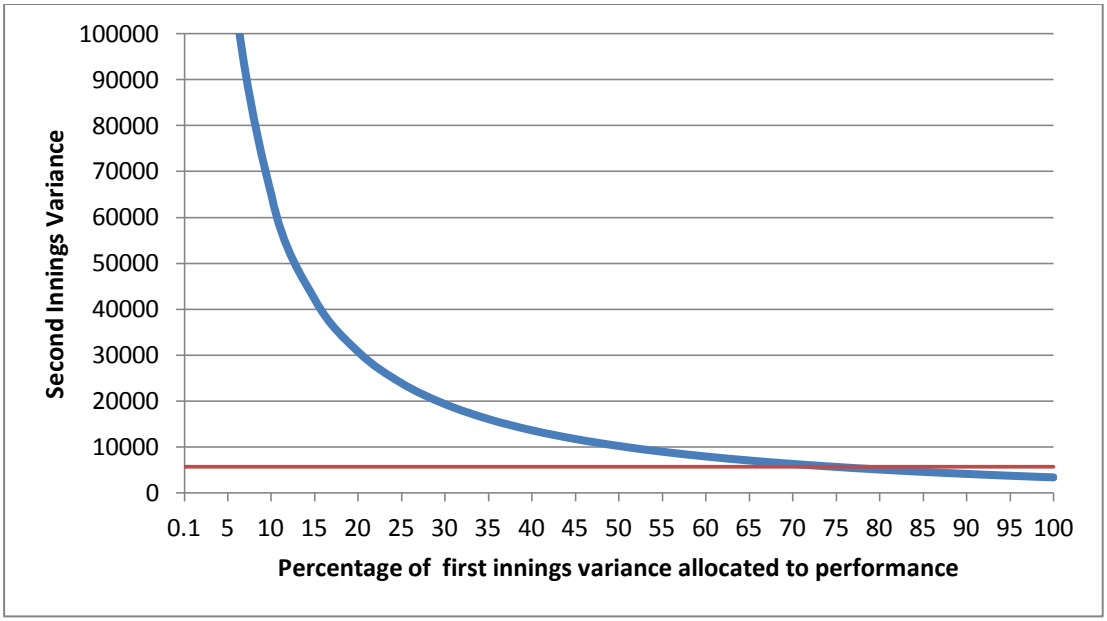
parameters α_δ and β_δ. In Figure A.3 we show the 21 means of the second-innings distribution, obtained from the 21 different variance splits. The mean is equal to the theoretical value of 243.3 in all cases, indicating that, in the absence of a second-innings advantage, the mean of the second-innings distribution is invariant to the chosen split of the first-innings variance.

Figure A.3: Second-innings mean for each split of first-innings variance



In Figure A.4 we show the variances of the second-innings distributions obtained from the 21 different splits. In order to show the data in a meaningful way we set the maximum point of the vertical axis to 100000 as the second-innings variance tends to ∞ as we allocate all the variance to conditions and none to performance. The red line is the observed second-innings variance of 5673.117 while the blue line is the variance of S_2 implied by split δ. We see that allocating approximately 75% of the first-innings variance to performance and 25% to conditions we get the closest second-innings variance to our observed value.

Figure A.4: Second-innings variance for each split of first-innings variance



The process that we have modelled is a rather complicated one and we were unable to derive the formulae for var(ρ) and var(χ) analytically. However, by investigating the results of the large number of splits of the first-innings variance that we used to determine Figure A.4, we were able to determine that the variance of the second-innings mean, satisfies the following equation

$$\sigma^2_{S_2,\delta} = \sigma^2_S\left(\frac{2}{\delta} - 1\right)$$

Given that we know $\sigma^2_{S_2}$ and σ^2_S, we can define the percentage of score variance that should be allocated to performance as

$$\delta = 100\frac{2\sigma^2_S}{\sigma^2_S + \sigma^2_{S_2}}$$

Appendix B: Numerical investigations into the second-innings advantage

In order to find the value for A_ρ that implies a second-innings advantage in scores of 4.7, we first investigate the effect on A_S of trying different values of A_ρ, assuming our estimated split of performance and conditions. We include negative second-innings advantages and determine the implied A_S for $A_\rho \in \{-10,-9,...,9,10\}$. The results are shown in Table B.1.

Table B.1: Values of A_S implied by different values of A_ρ

A_ρ	α	β	A_S	$\frac{A_S}{A_\rho}$
-10	-3.0533	0.013277	-13.312	1.331
-9	-3.07097	0.013277	-11.981	1.331
-8	-3.08865	0.013277	-10.65	1.331
-7	-3.10632	0.013277	-9.319	1.331
-6	-3.124	0.013277	-7.987	1.331
-5	-3.14167	0.013277	-6.656	1.331
-4	-3.15934	0.013277	-5.325	1.331
-3	-3.17702	0.013277	-3.994	1.331
-2	-3.19469	0.013277	-2.662	1.331
-1	-3.21237	0.013277	-1.331	1.331
0	-3.23004	0.013277	0	1.331
1	-3.24772	0.013277	1.331	1.331
2	-3.26539	0.013277	2.662	1.331
3	-3.28306	0.013277	3.994	1.331
4	-3.30074	0.013277	5.325	1.331
5	-3.31841	0.013277	6.656	1.331
6	-3.33609	0.013277	7.987	1.331
7	-3.35376	0.013277	9.319	1.331
8	-3.37144	0.013277	10.65	1.331
9	-3.38911	0.013277	11.981	1.331
10	-3.40678	0.013277	13.312	1.331

There are two interesting pieces of information that we can obtain from Table B.1. First, the value of β is constant while α is changing with the level of A_ρ. This indicates that A_p affects the mean of the second-innings distribution but not the variance. Second, we see that the ratio of A_S to A_ρ is constant. This means that an increase in the second-innings performance advantage increases the second-innings score advantage by a constant percentage. In fact, this ratio can be defined as follows.

$$\frac{A_S}{A_\rho} = \frac{\sigma_S^2}{\sigma_\rho^2}$$

The unobserved performance advantage can be derived from the other three variables, for which we have already estimated values.

$$A_\rho = \frac{A_S \sigma_\rho^2}{\sigma_S^2} = \frac{4.694 \times 2563.412}{3412.488} = 3.526$$

The implied relationship between the second-innings mean, the first-innings mean and the performance advantage is

$$\mu_{S_2} = \mu_S + \frac{\sigma_S^2}{\sigma_\rho^2} A_\rho$$

$$\mu_{S_2} = \mu_S + \frac{A_\rho}{\delta} \tag{29}$$

From Equation (29) we can determine the performance advantage implied any advantage in score.

Appendix C: Regression coefficients in the first-innings dynamic programmes

Table C.1: Runs model coefficients for "with restrictions" model

	With conditions		Without conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept: r=6	-3.4257	<0.0001	-3.0692	<0.0001
Intercept: r=4	-2.1626	<0.0001	-1.8073	<0.0001
Intercept: r=3	-2.0983	<0.0001	-1.7430	<0.0001
Intercept: r=2	-1.8887	<0.0001	-1.5336	<0.0001
Intercept: r=1	-1.2671	<0.0001	-0.9126	0.0003
W_0	0.2461	0.3262	0.2709	0.2800
W_1	0.1356	0.5861	0.1548	0.5344
W_2	-0.0837	0.7388	-0.0735	0.7696
W_3	-0.0604	0.8176	-0.0581	0.8243
i	0.00573	0.0479	0.00574	0.0478
i_2	-0.00513	<0.0001	-0.00488	<0.0001
$W_0 \times i$	0.00271	0.3309	0.00283	0.3313
$W_1 \times i$	0.00233	0.3915	0.00240	0.3915
$W_2 \times i$	0.00295	0.2825	0.00304	0.2828
$W_3 \times i$	0.00182	0.5258	0.00190	0.5258
χ	0.00155	<0.0001	N/A	N/A

Table C.2: Runs model coefficients for "without restrictions" model

	With conditions		Without conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept: r=6	-3.1692	<0.0001	-2.8525	<0.0001
Intercept: r=4	-2.3139	<0.0001	-1.9983	<0.0001
Intercept: r=3	-2.2658	<0.0001	-1.9502	<0.0001
Intercept: r=2	-1.8965	<0.0001	-1.5812	<0.0001
Intercept: r=1	-0.6865	<0.0001	-0.3719	<0.0001
pp	0.0137	0.1735	0.0291	0.0032
W_1	-0.0963	0.0182	-0.1018	0.0126
W_2	-0.1686	<0.0001	-0.1774	<0.0001
W_3	-0.2770	<0.0001	-0.2897	<0.0001
W_4	-0.4848	<0.0001	-0.5140	<0.0001
W_5	-0.5349	<0.0001	-0.5745	<0.0001
W_6	-0.3365	0.0077	-0.3698	0.0034
W_7	-0.8653	<0.0001	-0.9319	<0.0001
W_8	0.3773	0.3403	0.3261	0.4095
W_9	0.7335	0.1297	0.6785	0.1609
i	0.00392	<0.0001	0.00401	<0.0001
i_3	0.00665	0.0006	0.00673	0.0005
$W_3 \times i$	0.00010	0.7777	0.00009	0.8047
$W_4 \times i$	0.00061	0.1084	0.00066	0.0804
$W_5 \times i$	0.00012	0.7948	0.00018	0.7038
$W_6 \times i$	-0.00130	0.0326	-0.00129	0.0343
$W_7 \times i$	0.00055	0.5506	0.00067	0.4663
$W_8 \times i$	-0.00533	0.0026	-0.00530	0.0028
$W_9 \times i$	-0.00793	0.0004	-0.00794	0.0004
$W_3 \times i_3$	0.00204	0.4155	0.00198	0.4312
$W_4 \times i_3$	0.00207	0.3440	0.00198	0.3652
$W_5 \times i_3$	0.00624	0.0056	0.00627	0.0053
$W_6 \times i_3$	0.00683	0.0035	0.00685	0.0034
$W_7 \times i_3$	0.00530	0.0413	0.00531	0.0409
$W_8 \times i_3$	0.0129	<0.0001	0.0131	<0.0001
$W_9 \times i_3$	0.0162	<0.0001	0.0165	<0.0001
χ	0.00130	<0.0001	N/A	N/A

Table C.3: Wickets model coefficients for "with restrictions" model

	With Conditions		Without Conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept	-1.5467	<0.0001	-2.0799	<0.0001
W_0	0.1178	0.2718	0.0645	0.5440
W_1	-0.0149	0.8861	-0.0544	0.5992
W_2	-0.0797	0.4486	-0.1072	0.3064
W_3	-0.0903	0.4185	-0.1028	0.3561
i	0.00267	0.0544	0.00244	0.0773
i_2	-0.00276	0.1804	-0.00305	0.1384
χ	-0.00242	<0.0001	N/A	N/A

Table C.4: Wickets model coefficients for "without restrictions" model

	With Conditions		Without Conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept	-1.0443	0.0052	-2.2046	<0.0001
W_1	-0.0756	0.4560	-0.0649	0.5216
W_2	-0.0806	0.4137	-0.0588	0.5496
W_3	-0.1488	0.1292	-0.1165	0.2327
W_4	-0.2285	0.0212	-0.1858	0.0594
W_5	-0.1820	0.0700	-0.1278	0.2003
W_6	-0.1788	0.0798	-0.1207	0.2335
W_7	-0.2329	0.0254	-0.1691	0.1021
W_8	-0.1325	0.2193	-0.0661	0.5372
W_9	-0.0868	0.4446	-0.0132	0.9063
i	-0.00171	0.2834	0.0014	<0.0001
i_3	0.0102	<0.0001	0.0102	<0.0001
χ	-0.00483	0.0012	N/A	N/A
$\chi \times i$	0.000014	0.0354	N/A	N/A

Table C.5: Pr(Wide or No-ball) combined model

	With Conditions		Without Conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept	-1.8097	<0.0001	-1.6905	<0.0001
i	-0.00130	<0.0001	-0.0013	<0.0001
pp	-0.0234	0.1679	N/A	N/A
χ	0.00053	0.0699	N/A	N/A

Table C.6: Expected runs from a wide or no-ball combined model

	With Conditions		Without Conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept: $\tau_{ij} = 7$	-3.8262	<0.0001	-3.0237	<0.0001
Intercept: $\tau_{ij} = 5$	-2.7802	<0.0001	-1.9871	<0.0001
Intercept: $\tau_{ij} = 4$	-2.7245	<0.0001	-1.9317	<0.0001
Intercept: $\tau_{ij} = 3$	-2.5464	<0.0001	-1.7544	<0.0001
Intercept: $\tau_{ij} = 2$	-2.0137	<0.0001	-1.2239	<0.0001
i	0.00231	<0.0001	0.00236	<0.0001
pp	-0.0129	0.8167	N/A	N/A
χ	0.00326	0.0005	N/A	N/A

Appendix D: Regression coefficients in the second-innings dynamic programmes

Table D.1: Runs model coefficients for "with restrictions" model

	With conditions		Without conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept: r=6	-3.3988	<0.0001	-3.2615	<0.0001
Intercept: r=4	-2.1313	<0.0001	-1.9944	<0.0001
Intercept: r=3	-2.0770	<0.0001	-1.9401	<0.0001
Intercept: r=2	-1.8701	<0.0001	-1.7332	<0.0001
Intercept: r=1	-1.2635	<0.0001	-1.1268	<0.0001
pp	-0.00310	0.9158	0.00429	0.8833
$i \times pp$	0.00072	0.1616	0.00072	0.1601
W_0	0.3772	<0.0001	0.3889	<0.0001
W_1	0.2507	<0.0001	0.2595	<0.0001
W_2	0.1514	<0.0001	0.1575	<0.0001
W_3	0.0512	0.2344	0.0553	0.1989
i	0.00676	<0.0001	0.00682	<0.0001
i_2	-0.00555	<0.0001	-0.00557	<0.0001
k^*	0.1762	<0.0001	0.2174	<0.0001
χ	0.00076	0.0063	N/A	N/A

Table D.2: Runs model coefficients for "without restrictions" model

	With conditions		Without conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept: r=6	-3.3179	<0.0001	-3.0410	<0.0001
Intercept: r=4	-2.4316	<0.0001	-2.1554	<0.0001
Intercept: r=3	-2.3776	<0.0001	-2.1015	<0.0001
Intercept: r=2	-2.0210	<0.0001	-1.7451	<0.0001
Intercept: r=1	-0.8914	<0.0001	-0.6160	<0.0001
pp	-0.0396	0.7358	-0.0290	0.8047
$i \times pp$	0.000386	0.5997	0.000368	0.6165
$i_3 \times pp$	-0.00137	0.3183	-0.00133	0.3352
$i_4 \times pp$	0.00238	0.4106	0.00227	0.4328
$i_5 \times pp$	0.00577	0.4284	0.00627	0.3898
W_3	-0.0698	<0.0001	-0.0735	<0.0001
W_4	-0.1565	<0.0001	-0.1654	<0.0001
W_5	-0.2431	<0.0001	-0.2554	<0.0001
W_6	-0.2395	<0.0001	-0.2554	<0.0001
W_7	-0.4303	<0.0001	-0.4533	<0.0001
W_8	-0.4987	<0.0001	-0.5265	<0.0001
W_9	-0.6445	<0.0001	-0.6832	<0.0001
i	0.00206	<0.0001	0.00211	<0.0001
i_3	0.00122	0.0877	0.00120	0.0943
i_4	0.00724	<0.0001	0.00734	<0.0001
i_5	0.00377	0.3982	0.00408	0.3605
k^*	0.3386	<0.0001	0.3630	<0.0001
$k^* \times pp$	0.00670	0.8703	0.0150	0.7139
χ	0.00124	<0.0001	N/A	N/A

Table D.3: Wickets model coefficients for "with restrictions" model

	With Conditions		Without Conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept	-2.0473	<0.0001	-2.2736	<0.0001
pp	0.0463	0.4688	0.0341	0.5917
W_0	0.1366	0.1487	0.1177	0.2109
W_1	0.1712	0.0628	0.1572	0.0864
W_2	0.0548	0.5547	0.0452	0.6257
W_3	0.0772	0.4368	0.0713	0.4724
i	-0.00031	0.8263	-0.00041	0.7677
i_2	0.00166	0.4520	0.00169	0.4436
k^*	0.2689	0.0008	0.1999	0.0060
$i \times pp$	-0.00054	0.6358	-0.00054	0.6359
χ	-0.00126	0.0433	N/A	N/A

Table D.4: Wickets model coefficients for "with restrictions" model

	With Conditions		Without Conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept	-1.8010	<0.0001	-2.1610	<0.0001
pp	-0.4024	0.1537	-0.4188	0.1377
$i \times pp$	0.00206	0.2424	0.00210	0.2336
W_3	-0.0630	0.1490	-0.0583	0.1810
W_4	-0.00536	0.9042	0.00610	0.8907
W_5	0.00304	0.9499	0.0195	0.6856
W_6	0.00252	0.0527	0.0457	0.3820
W_7	0.1520	0.0117	0.1810	0.0024
W_8	-0.00708	0.9174	0.0285	0.6712
W_9	0.1013	0.2235	0.1510	0.0642
i	-0.00029	0.7189	-0.00036	0.6534
i_3	0.00377	0.0263	0.00382	0.0242
i_4	-0.00004	0.9916	-0.00012	0.9727
i_5	0.0188	0.0157	0.0182	0.0190
$i_3 \times pp$	-0.00644	0.0441	-0.00651	0.0418
$i_4 \times pp$	0.00570	0.3459	0.00581	0.3358
$i_5 \times pp$	-0.00670	0.6101	-0.00708	0.5889
k^*	0.1796	0.0021	0.1504	0.00941
$k^* \times pp$	0.1875	0.0375	0.1771	0.0489
χ	-0.00162	0.0013	N/A	N/A

Table D.5: Pr(Wide or No-ball) combined model

	With Conditions		Without Conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept	-1.6666	<0.0001	-1.6567	<0.0001
i	-0.00156	<0.0001	-0.00156	<0.0001
pp	-0.0757	<0.0001	-0.0752	<0.0001
χ	0.000041	0.9020	N/A	N/A

Table D.6: Runs from a wide or no-ball combined model

	With Conditions		Without Conditions	
Variable	**Coefficient**	**P-value**	**Coefficient**	**P-value**
Intercept: $\tau_{ij} = 7$	-3.3654	<0.0001	-3.0115	<0.0001
Intercept: $\tau_{ij} = 5$	-2.1749	<0.0001	-1.8231	<0.0001
Intercept: $\tau_{ij} = 4$	-2.1440	<0.0001	-1.7923	<0.0001
Intercept: $\tau_{ij} = 3$	-1.9131	<0.0001	-1.5615	<0.0001
Intercept: $\tau_{ij} = 2$	-1.4427	<0.0001	-1.0914	<0.0001
i	0.00112	0.0049	0.00108	0.0068
pp	-0.0324	0.6026	-0.0116	0.8480
k^*	0.0645	0.5746	0.1145	0.2963
χ	0.00163	0.1439	N/A	N/A